DISCLAIMER NOTICE

THIS DOCUMENT IS THE BEST QUALITY AVAILABLE.

COPY FURNISHED CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

ARPA Order No. 73-59

Unclassified Extracts

from

ARPA SEMIANNUAL TECHNICAL NOTE

APPLICATION OF RECOGNITION THEORY

TO DETECTION, DISCRIMINATION, AND DECISION

Contract No. AF30(602)-2112

Project No. 4983 Task No. 55147

Melpar Job No. A1066.01

Period Covered: 1 October 1959 - 30 June 1960

Prepared for:

ROME AIR DEVELOPMENT CENTER

AIR RESEARCH AND DEVELOPMENT COMMAND

United States Air Force

Griffiss Air Force Base New York

The concept playing a central role in the theory which will be described is the notion that the ensemble of points in signal space which represents a set of nonidentical events belonging to a common category must be close to each other as measured by some as yet unknown method of measuring distance, since the points represent events which are close to each other in the sense that they are members of the same category. Mathematically speaking, the fundamental notion underlying the theory is that similarity (closeness in the sense of belonging to the same class or category) is expressible by a metric (a method of measuring distance) by which points representing examples of the category we wish to recognize are found to lie close to each other.

To give credence to this conjecture, consider what we mean by the abstract concept of a class. According to one of the possible definitions, a class is a collection of things which have some common properties. By a modification of this thought, a class could be characterized by the common properties of its members. A metric by which points representing examples of a class are close to each other must therefore operate chiefly on the common properties of the examples and must ignore, to a large extent, those properties not present in each example. As a consequence of this argument, if a metric were found which called examples of the class close, somehow it must exhibit their common properties.

To present this fundamental idea in a slightly different way, we can state that a transformation on the signal space which is capable of clustering the points representing the examples of the class must operate primarily on the common properties of the examples. A simple illustration

of this idea is shown in Figure 1, where the ensemble of points is spread out in signal space (only a two-dimensional space is shown for ease of illustration) but a transformation T of the space is able to cluster the points of the ensemble.

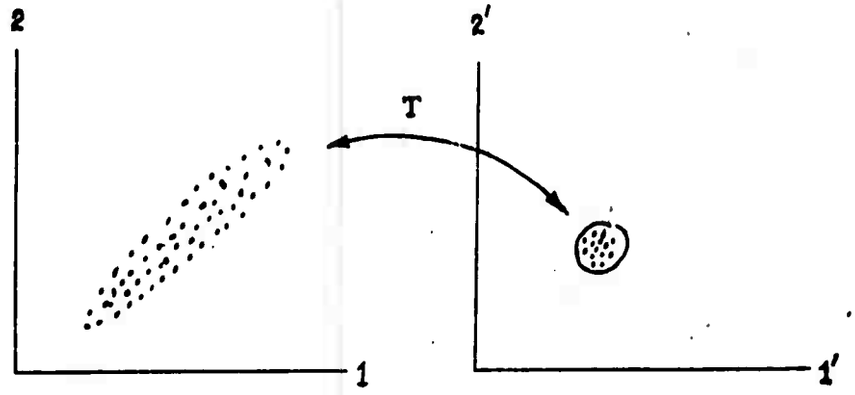


Figure 1. Clustering by Transformation

In the above example neither the signal's property represented by coordinate 1 nor that represented by coordinate 2 is sufficient to describe the class, for the spread in both is large over the ensemble of points. Some function of the two coordinates on the other hand, would exhibit the common property that the ratio of the value of coordinate 2 to that of coordinate 1 in each point in the ensemble is nearly unity. In this specific instance, of course, simple correlation between the two coordinates would exhibit this property, but in more general situations simple correlation will not suffice.

If the signal space shown in Figure 1 were flexible (as if made of a rubber sheet), the transformation T would express the manner in which various portions of the space must be stretched or compressed, in order to bring the points together most closely.

Although thinking of transformations of the space is not as general as thinking about exotic ways of measuring "distance" in the original space, the former is a rigorously correct and easily visualized analogy for many important classes of metrics.

Mathematical techniques have been developed to automatically find the "best" metric of "best" transformation of given classes of metrics according to suitable criteria which establish "best".

As any mathematical theory, the one which evolved from the preceeding ideas is based on certain assumptions. The most basic assumption is that the N-dimensional signal space representation of events exemplifying their respective classes is complete enough to contain information about the common properties which serve to characterize the classes. The significance of this assumption is appreciated if we consider, for example, that the signal space contains all the information that a black and white television picture could present of the physical objects making up the sequence of events which constitute the examples of a class. No matter how ingenious the data processing schemes that we might evolve are, objects belonging to the category "red things" could not be identified, because representation of the examples by black and white television simply does not contain color information. For any practical situation one must rely on engineering judgment and intuition to determine if the model of the real world (the signal space) is complete enough. Fortunately, in most cases, this determination may be made with considerable confidence.

A second assumption states the class of transformations or the class of metrics within which we look for the "best". This assumption

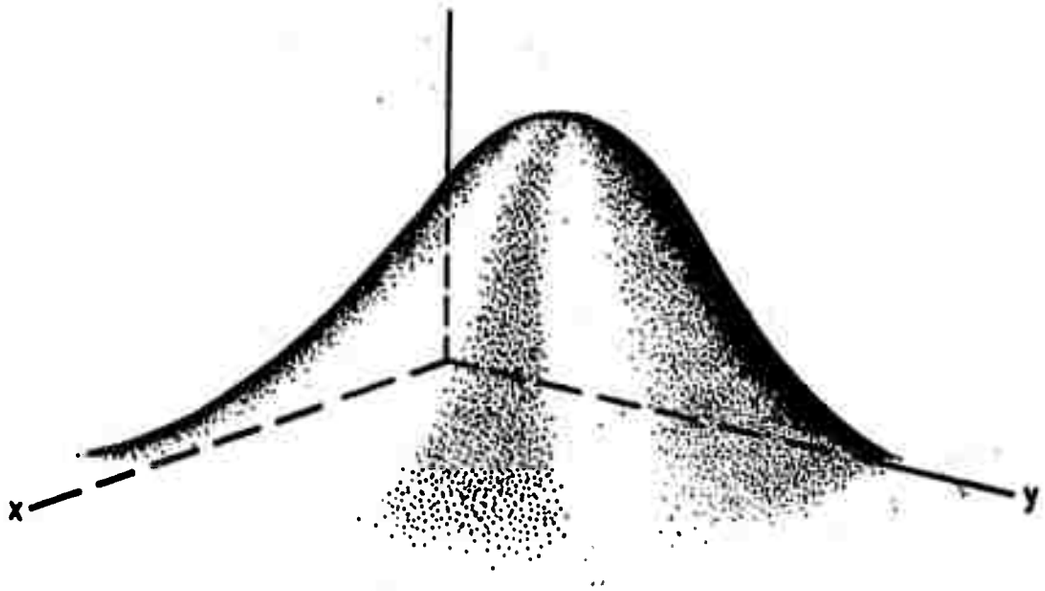
2. A SPECIAL THEORY OF SIMILARITY

2.1 Similarity

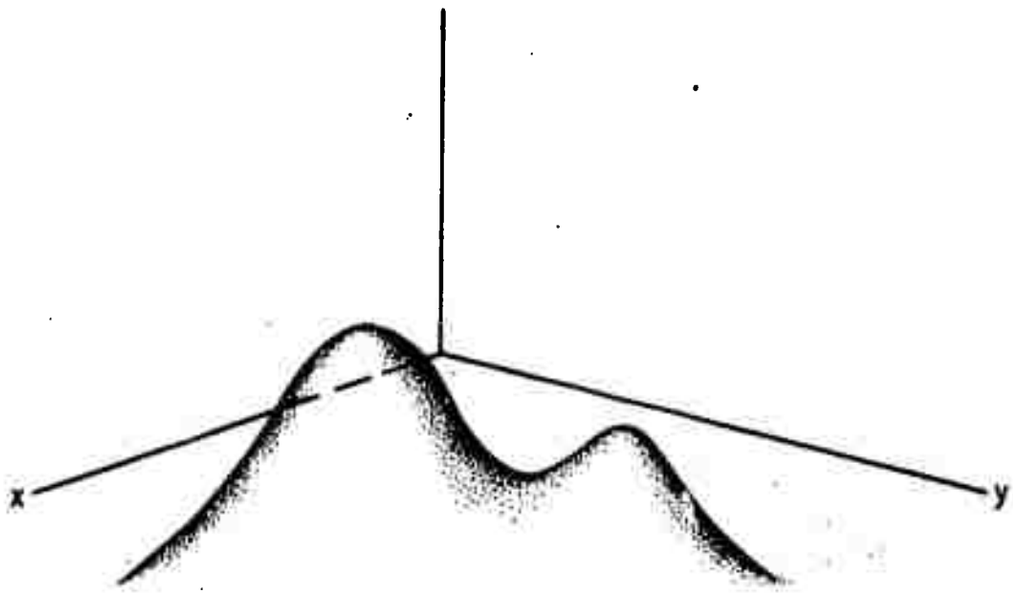
The central problem of pattern recognition is viewed in this work as the problem of developing a function of a point and a set of points in an N-dimensional space to partition the space into a number of regions corresponding to the categories to which the known set of points belong. A convenient special—but not essential—way of thinking about this partitioning function is to consider it formed from a set of functions, one for each category, where each function measures the "likelihood"* with which an arbitrary point of the space could best fit into the particular function's own category. In a sense, each function measures the similarity of an arbitrary point of the space to a category and the partitioning function assigns the arbitrary point to that category to which the point is most similar.

The foregoing concept of partitioning the signal space is illustrated in Figure 2 where the signal space has two dimensions and the space is to be partitioned into two categories. In Figure 2a, the height of the surface above the x-y plane expresses the likelihood that a point belongs to Category 1, while that of the surface in Figure 2b expresses the likelihood that the point belongs to Category 2. The intersection between the two surfaces, shown in Figure 3a and b, marks the boundary between Region 1 where points are more likely to belong to Category 1 than to Category 2, and Region 2, where the reverse is true.

* Although the term "likelihood" has an already well-defined meaning in decision theory, it is used here in a qualitative way to emphasize the similarity between fundamental ideas in decision theory and in the theory which is here described.

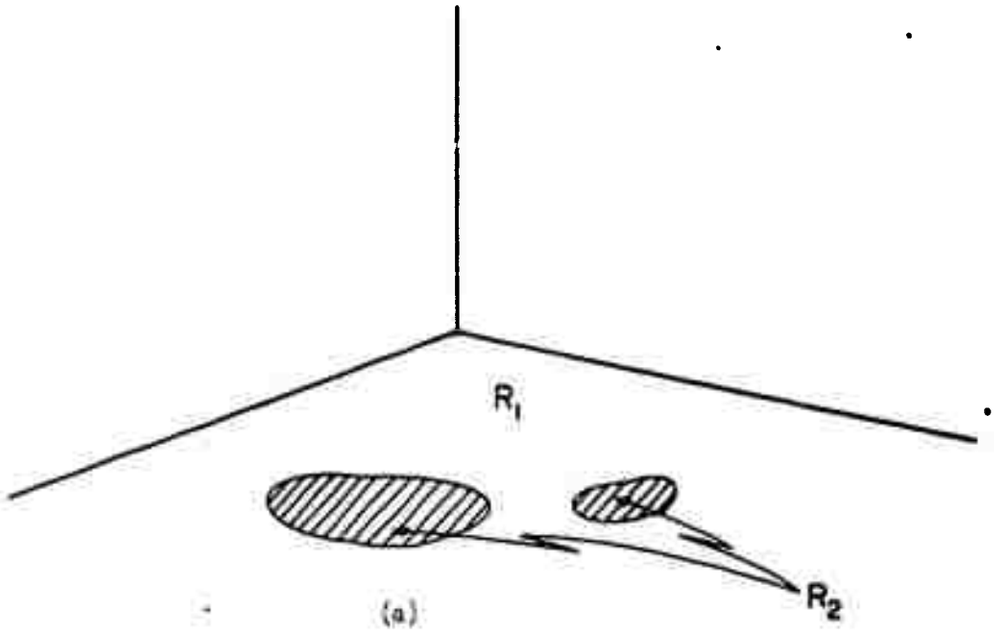


a) "Likelihood" of Membership in Category 1



b) "Likelihood" of Membership in Category 2

Figure 2. Likelihood of Membership in Two Categories



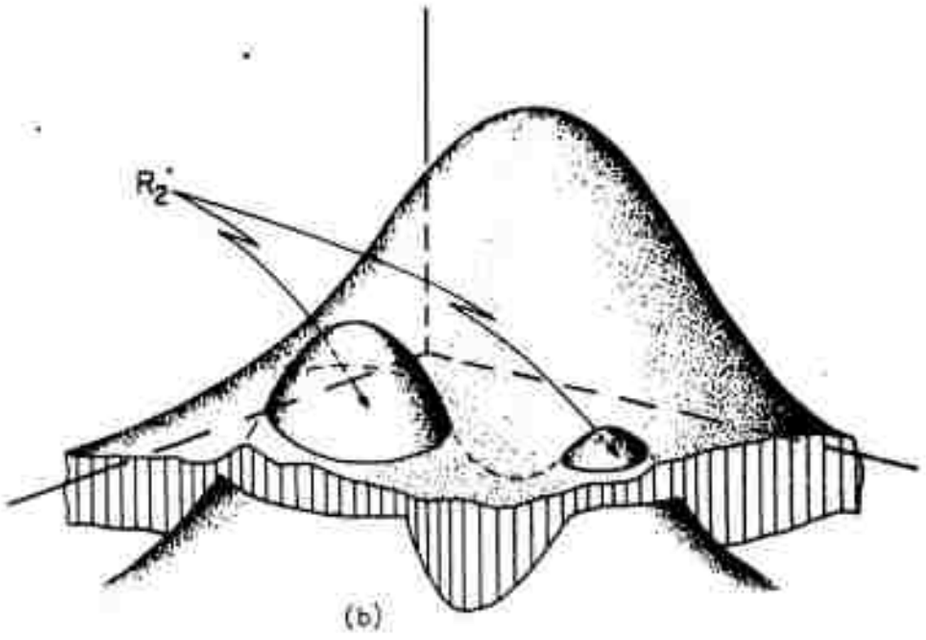


Figure 3. Classification by Maximum Likelihood Ratio

For each category of interest a set of likelihood ratios may be computed which express the relative likelihood that a point in question belongs to the category of interest rather than to any of the others. From the maximum of all likelihood ratios which correspond to a given point, we may infer to which category the point most likely belongs.

The reader will recognize the idea of making decisions based on the maximum likelihood ratio as one of the important concepts of decision theory. The objective of the preceding discourse is, therefore, simply to make the statement that once a function measuring the likelihood that a point belongs to a given category is developed, there is at least one well-established precedent for partitioning signal space into regions which are associated with the different categories. The resulting regions are like a template which serves to categorize points depending upon whether they are covered or are left uncovered by the template. Although in the rest of this chapter partitioning the signal space is based on a measure of similarity which resembles the likelihood ratio only in the manner in which it is used, it is shown elsewhere that, in certain cases, decisions based on the measure of similarity are identical to those based on the maximum likelihood ratio.

One might wonder whether the error criterion by which similarity to a class of things is measured should be based on known members of the class only, or also on the additional knowledge gained from a set of things which do not belong to the class. The philosophical question posed by these two possibilities is whether one is aided in learning to recognize membership in a category if, during the period of learning, examples of nonmembers of the category are also given. It seems plausible that increasing the knowledge available on members and nonmembers of the category may better the separation

between categories. There are significant categories, however, where knowledge of nonmembers does not help to determine what constitutes membership in the category. The analogous situation in decision theory is pointed out later.

In the first three chapters of this report, a quantitative measure of similarity is developed in a special theory where similarity is considered as a property of only the point to be compared and the set of points which belong to the category to be learned. In later chapters, however, methods will be discussed for letting known nonmembers of the class influence the development of measures of similarity.

In the special theory of the first three chapters, similarity of an event P to a category is measured by the closeness of P to every one of those events $\{F_m\}$ known to be contained in the category. Similarity S is regarded as the average "distance" between P and the class of events represented by the set $\{F_m\}$ of its examples.

Two things should be noted about the foregoing definition of similarity. One is that the method of measuring distance does not influence the definition. Indeed, distance is not meant here in the ordinary Euclidean sense; it may mean "closeness" in some arbitrary, abstract property of the set $\{F_m\}$ which has yet to be determined. The second thing to note is that the concept of distance between points, or distance in general, is not fundamental to a concept of similarity. The only aspect of similarity really considered essential is that it is a real valued function of a point and a set which allows the ordering of points according to their similarity to the set. The concept of distance is introduced here as a mathematical convenience based on intuitive notions

of similarity. It will be apparent later how this forms part of the assumptions stated in the Introduction as underlying the theory to be presented. Even with the introduction of the concept of distance there are other ways of defining similarity. Nearness to the closest member of the set is one such possibility. This implies that an event is similar to a class of events if it is close in some sense to any member of the class. It is not the purpose of this chapter to philosophize about the relative merits of these different ways of defining similarity. Their advantages and disadvantages will become apparent as this theory is developed, and the reader will be able to judge for himself which set of assumptions is most applicable under a given set of circumstances.

To summarize the foregoing remarks, for the purposes of the special theory, similarity $S\left(P, \{F_m\}\right)$ of a point P and a set of points $\{F_m\}$ exemplifying a class will be defined as the average distance between the point P and the M members of the set $\{F_m\}$. This definition is expressed by Equation 2.1, where the metric d( )—the method of measuring distance between two points—is left unspecified.

$$S\left(P, \{F_m\}\right) = \frac{1}{M} \sum_{m=1}^{M} d(P, F_m). \qquad (2.1)$$

To deserve the name metric, the function d( ) must satisfy the usual conditions stated in Equation 2.2 a, b, c and d.

$$d(A,B) = d(B,A) \qquad \text{(symmetric function)} \qquad (2.2a)$$

$$d(A,C) \leq d(A,B) + d(B,C) \qquad \text{(triangle inequality)} \qquad (2.2b)$$

$$d(A,B) \geq 0 \qquad \text{(non-negative)} \qquad (2.2c)$$

$$d(A,B) = 0 \text{ if, and only if, } A = B \qquad (2.2d)$$

2.2 Optimization and Feature Weighting

In the definition of similarity of the preceding section the average distance between a point and a set of points served to measure similarity of a point to a set. The method of measuring distance, however, was left unspecified and was understood to refer to distance in perhaps some abstract property of the set. In this section the criteria for finding the "best" choice of the metric are discussed, and this optimization is applied to a specific and simple class of metrics which has interesting and useful properties.

Useful notions of "best" in mathematics are often associated with finding the extrema of the functional to be optimized. We may seek to minimize the average cost of our decisions or we may maximize the probability of estimating correctly the value of a random variable. In the problem above, a useful metric, optimal in one sense, is one which minimizes the average distance between members of the same set subject to certain suitable constraints devised to assure a nontrivial solution. If the metric is thought of as extracting that property of the set in which like events are clustered, then the average distance between members of the set is a measure of the size of the cluster so formed. Minimization of the average distance is then a choice of a metric which minimizes the size of the cluster

and therefore extracts that property of the set in which they are most alike. It is only proper that a distance measure shall minimize the average distance between those events which are selected to exemplify events that are "close".

Although this preceding criterion for finding the best solution is a very reasonable and meaningful assumption on which to base the special theory, it is by no means the only possibility. Minimization of the maximum distance between members of a set is just one of the possible alternatives that immediately suggests itself. It should be pointed out that ultimately the best solution is that which results in the largest number of correct classifications of events. Making the largest number of correct decisions on the known events is thus to be maximized and is itself a suitable criterion of optimization which will be dealt with elsewhere in this report. Since the primary purpose of this chapter is to outline a point of view regarding pattern recognition through a special example, the choice of "best" previously described and stated in Equation 2.3 will be used, for it leads to very useful solutions with relative simplicity of the mathematics involved. In Equation 2.3 F_p and F_m are the p^{th} and m^{th} members of the set $\{F_m\}$.

$$\min \overline{d(F_p,F_m)}^{p,m} = \min \frac{1}{M(M-1)} \left[\sum_{m=1}^{M} \sum_{p=1}^{M} d(F_p,F_m) \right], \text{ over all choices of } d(\ ). \tag{2.3}$$

Of the many different mathematical forms which a metric may take, in the special theory here described only metrics of the form given by Equation 2.4 will be considered. The intuitive notions underlying

the choice of the metric in this form are based on ideas of "feature weighting" which will be developed below.

$$d(A,B) = \sqrt{\sum_{n=1}^{N} W_n^2 (a_n - b_n)^2}. \tag{2.4}$$

In the familiar Euclidean N-dimensional space the distance between the two points A and B is defined by Equation 2.5. If A and B are expressed in terms of an orthonormal coordinate system $\{\theta_n\}$, then d(A,B) of Equation 2.5 can be written as in Equation 2.6, where a_n and b_n, respectively, are the coordinates of A and B in the direction of θ_n.

$$d(A,B) = |A - B|. \tag{2.5}$$

$$d(A,B) = \sqrt{\sum_{n=1}^{N} (a_n - b_n)^2}. \tag{2.6}$$

We must realize, of course, that the features of the events represented by the different coordinate directions θ_n are not all equally important in influencing the definition of the category to which like events belong. Therefore it is reasonable that in comparing two points feature by feature (as is expressed in Equation 2.6), features with decreasing significance should be weighted with decreasing weights W_n. The idea of feature weighting is expressed by a metric somewhat more general than the conventional Euclidean metric. The modification is given in Equation 2.7, where W_n is the feature weighting coefficient.

$$d(A,B) = \sqrt{\sum_{n=1}^{N} \left[W_n \, (a_n - b_n) \right]^2}. \qquad (2.7)$$

It is readily verified that the above metric satisfies the conditions stated in Equation 2.2 if none of the W_n's is zero; if any of the W_n coefficients is zero, Equation 2.2d is not satisfied.

It is important to note that the above metric gives a numerical measure of "closeness" between two points, A and B, which is strongly influenced by the particular set of similar events $\{F_m\}$. This is a logical result, for a measure of similarity between A and B should depend on how our notions of similarity were shaped by the set of events known to be similar. When we deal with a different set of events which have different similar features, our judgement of similarity between A and B will also be based on finding agreement between them along a changed set of their features.

An alternate and instructive way of explaining the significance of the class of metrics given in Equation 2.4 is to recall the analogy made in the Introduction regarding transformations of the signal space. There, the problem of expressing what was similar among a set of events of the same category was accomplished by finding that transformation of the signal space (again, subject to suitable constraints), which will cluster most highly the transformed events in the new space. If we restrict ourselves to those linear transformations of the signal space which involve only scale factor changes of the coordinates and if we measure distance

in the new space by the Euclidean metric, then the Euclidean distance between two points after their linear transformation is equivalent to the feature weighting metric of Equation 2.4. This equivalence is shown below, where A' and B' are vectors obtained from A and B by a linear transformation. The most general linear transformation is expressed by Equation 2.9, where a'_n is the n^{th} coordinate of the transformed vector A and b'_n is that of the vector B.

$$A = \sum_{n=1}^{N} a_n \theta_n \qquad \text{and} \qquad B = \sum_{n=1}^{N} b_n \theta_n \tag{2.8a}$$

$$[A'] = [A][W] \qquad [B'] = [B][W]. \tag{2.8b}$$

$$[A' - B'] = [A - B][W] \tag{2.8c}$$

$$\left[(a'_1-b'_1),(a'_2-b'_2),\ldots,(a'_N-b'_N)\right] = \left[(a_1-b_1),(a_2-b_2),\ldots,(a_N-b_N)\right]\begin{bmatrix} w_{11}w_{12}\cdots w_{1N} \\ w_{21}w_{22}\cdots w_{2N} \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ w_{N1}w_{N2}\cdots w_{NN} \end{bmatrix} \tag{2.9}$$

The Euclidean distance between A' and B', $d_E(A', B')$, is given in Equation 2.10.

$$d_E(A',B') = \sqrt{\sum_{n=1}^{N} (a'_n - b'_n)^2} = \sqrt{\sum_{n=1}^{N} \left[\sum_{s=1}^{N} w_{ns}\,(a_s - b_s)\right]^2}\,. \tag{2.10}$$

If the linear transformation involves only scale factor changes of the coordinates, only the elements on the main diagonal of the W matrix are non-zero, thus reducing $d_E(A', B')$, in this special case, to the form given in Equation 2.11.

$$\text{Special } d_E(A', B') = \sqrt{\sum_{n=1}^{N} w_{nn}^2 (a_n - b_n)^2}. \tag{2.11}$$

The above class of metrics will be used in Equation 2.3 to minimize the average distance between the set of points. Because of the mathematical difficulty of minimizing the sum of square roots of quantities, we will minimize instead the mean-square distance when members of $\{F_m\}$ are compared with each other.

The mathematical formulation of the above minimization is given in Equations 2.12a and 2.12b. The significance of the constraint 2.12b is, for the case considered, that every weight w_{nn} is a number between 0 and 1 (w_{nn}'s turn out to be positive) and it can be interpreted as the fractional value of the features θ_n which they weight. w_{nn} denotes the fractional value which is assigned in the total measure of distance to the degree of agreement that exists between the components of the compared vectors.

$$\overline{D^2} = \frac{1}{M(M-1)} \sum_{p=1}^{M} \sum_{m=1}^{M} \sum_{n=1}^{N} w_{nn}^2 (f_{mn} - f_{pn})^2 = \text{minimum}, \tag{2.12a}$$

$$\text{if } \sum_{n=1}^{N} w_{nn} = 1. \tag{2.12b}$$

Although the constraint of 2.12b is appealing from a feature-weighting point of view, from a strictly mathematical standpoint it leaves much to be desired. It does not guarantee, for instance, that a simple shrinkage in the size of the signal space is disallowed. Such a shrinkage would not change the relative orientation of the points to each other, the property really requiring alteration. The constraint given in Equation 2.13, on the other hand, states that the volume of the space is constant as if the space were filled with an incompressible fluid. Here one merely wishes to determine what kind of a rectangular box could contain the space so as to minimize the mean-square distance among a set of points imbedded in the space.

$$\prod_{n=1}^{N} w_{nn} = 1. \tag{2.13}$$

The minimization problem with both of these constraints will be worked out in the following equations, and it will be seen that the results are quite similar.

Interchanging the order of summations and expanding the squared expression in Equation 2.12a yields Equation 2.14, where it is recognized that the factor multiplying w_{nn}^2 is the variance of the coefficients of the θ_n coordinate. Minimization of Equation 2.14 under the constraint 2.12b. yields Equation 2.15, where ρ is an arbitrary constant. Imposing contraint

2.12b again, we can solve for w_{nn}, obtaining Equation 2.14.

$$\overline{D^2} = \frac{M}{(M-1)} \sum_{n=1}^{N} w_{nn}^2 \left[\frac{1}{M} \sum_{m=1}^{M} f_{mn}^2 + \frac{1}{M} \sum_{p=1}^{M} f_{pn}^2 - 2\left(\frac{1}{M}\sum_{m=1}^{M} f_{mn}\right)\left(\frac{1}{M}\sum_{p=1}^{M} f_{pn}\right)\right] \quad (2.14)$$

$$\overline{D^2} = \frac{2M}{(M-1)} \sum_{n=1}^{N} w_{nn}^2 \left(\overline{f_n^2} - \overline{f_n}^2\right) = \frac{2M}{(M-1)} \sum_{n=1}^{N} w_{nn}^2 \sigma_n^2 .$$

$$\left[w_{nn} \sigma_n^2 - \rho\right] = 0, \qquad \text{for } n = 1, 2, \ldots, N. \quad (2.15)$$

$$w_{nn} = \frac{\rho}{\sigma_n^2} = \frac{1}{\sigma_n^2 \sum_{p=1}^{N} \frac{1}{\sigma_p^2}} . \quad (2.16)$$

That the values of w_{nn} so found are indeed those which minimize $\overline{D^2}$ of Equation 2.12a can be seen by noting that $\overline{D^2}$ is an elliptic paraboloid in an N-dimensional space and the constraint of 2.12b is a plane of the same dimensions. For a three-dimensional case, this is illustrated in Figure 4. The intersection of the elliptic paraboloid with the plane is a curve whose only point of zero derivative is a minimum.

The physical interpretation of weighting features by the reciprocal of their variances is given below.

If the variance of a coordinate of the ensemble is large, then the corresponding w_{nn} is small, indicating that small weight is to be given in the overall measure of distance to a feature of large variation. If the variance of the magnitude of a given coordinate θ_n is small, on the other hand, then its value can be accurately anticipated; therefore θ_n should be counted heavily in a measure of similarity. It is important to note

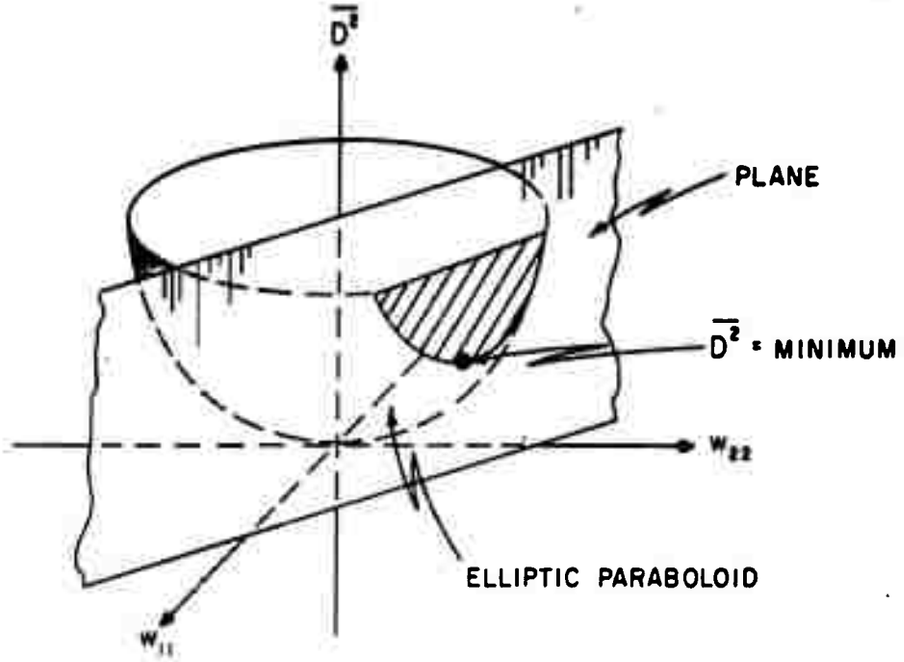


Figure 4. Geometric Interpretation of Minimization

that in the extreme case, where the variance of the magnitude of a component of the set is zero, the corresponding w_{nn} in Equation 2.16 is equal to unity with all other w_{nn}'s equal to zero. In this case, although Equation 2.11 is not a legitimate metric since it does not satisfy Equation 2.2, it is still a meaningful measure of similarity. If any coordinate occurs with identical magnitudes in all members of the set, then it is an "all important" feature of the set and nothing else needs to be considered in judging the events similar. Judging membership in a category by such an "all important" feature may, of course, result in the incorrect inclusion of nonmembers into the category. For instance "red, nearly circular figures" have the color red as a common attribute. The transformation described thus far would pick out "red" as an all important feature and would judge membership in the category of "red, nearly circular figures" only by the color of the compared object. A red square, for instance, would thus be misclassified and judged to be a "red, nearly circular figure". Given only examples of the category, on the other hand, such results would probably be expected. Later on, however, where labeled examples of all categories of interest are assumed given, only those attributes are emphasized in which members of a category are alike and in which they differ from those of other categories.

It should be noted that the weighting coefficients do not necessarily decrease monotonically in the above feature weighting which minimizes the mean-square distance among M given examples of the class. Furthermore, the results of Equation 2.16 or 2.18 are independent of the

particular orthonormal system of coordinates. Equations 2.16 and 2.18 simply state that the weighting coefficient is inversely proportional to the variance or to the standard deviation of the ensemble along the corresponding coordinate. The numerical values of the variances, on the other hand, do depend on the coordinate system.

If we use the mathematically more appealing constraint of Equation 2.13 in place of that in 2.12b, we obtain Equation 2.17.

$$\min \overline{D^2} = \min 2 \sum_{n=1}^{N} w_{nn}^2 \sigma_n^2 \quad \text{for} \prod_{n=1}^{N} w_{nn} = 1; \tag{2.17a}$$

$$\sum_{n=1}^{N} dw_{nn} \left[w_{nn} \sigma_n^2 - \lambda \prod_{k \neq n}^{N} w_{kk} \right] = 0. \tag{2.17b}$$

It is readily seen that by applying Equation 2.17a, the expression of 2.17b is equivalent to Equation 2.18a, where the bracketed expression must be zero for all values of n. This substitution leads to Equation 2.18b which may be reduced to Equation 2.18c by application of Equation 2.17a once more.

$$\sum_{n=1}^{N} dw_{nn} \left(w_{nn} \sigma_n^2 - \frac{\lambda}{w_{nn}} \right) = 0 \tag{2.18a}$$

$$w_{nn} = \frac{\sqrt{\lambda}}{\sigma_n} \tag{2.18b}$$

$$w_{nn} = \left(\prod_{p=1}^{N} \sigma_p \right)^{1/N} \frac{1}{\sigma_n} \qquad (2.18c)$$

Thus it is seen that the feature weighting coefficient w_{nn} is proportional to the reciprocal standard deviation of the n^{th} coordinates, thereby lending itself to the same kind of interpretation as before.

2.3 Describing the Category

The set of known members is the best description of the category. Following the practice of probability theory, this set of similar events can be described by its statistics; the ensemble mean, variance, and higher moments can be specified as its characteristic properties. For our purposes a more suitable description of our idea of the category, on the other hand, is found in the specific form of the function S of Equation 2.1 developed from the set of similar events to measure membership in the category. A marked disadvantage of S is that (in a machine which implements its application) the amount of storage capacity which must be available is proportional to the number of events introduced and is thus a growing quantity. For this reason a description of the set of points is desired in the form of a point E which may be considered most typical of the ensemble of points belonging to the set. Describing the category by means of a single point is analogous to designating a particular capital A as characterizing the set of different capital A's that are encountered. This single A takes the place of the entire ensemble of A's and represents it by being the typifying example of the set. The most important attribute to the typifying example, from the point of view of

correctly representing the set, is that the "distance" measured between an arbitrary point P and E should agree with the mean-square distance measured by the function S between P and members of the set. The distance in both cases is measured with the metric developed in the preceding section. The equality of these distances is stated in Equation 2.19, where e_n is the coordinate of E in the θ_n direction and p_n is the coordinate of P in the same direction.

$$S(P, \{F_m\}) = \frac{1}{K} \sum_{m=1}^{K} \sum_{n=1}^{N} w_n^2 (p_n - f_{mn})^2 = \sum_{n=1}^{N} w_n^2 (p_n - e_n)^2. \qquad (2.19)$$

Interchanging the order of summations, expanding the squares, and collecting like terms yields Equation 2.20.

$$\sum_{n=1}^{N} w_n^2 \left[e_n^2 - 2p_n e_n + (\overline{f_n^2} - 2p_n \overline{f_n}) \right] = 0. \qquad (2.20)$$

This equation does not have a unique solution unless further constraints are imposed. A convenient set of constraints is the requirement that the above equality hold for any choice of the metric. This can be shown to mean that the equation must hold for each n. Under this constraint the unique solution for E is given by Equation 2.21a and 2.21b.

$$E = \sum_{n=1}^{N} e_n \theta_n, \qquad (2.21a)$$

where

$$e_n = p_n \pm \sqrt{p_n^2 + \overline{f_n^2} - 2p_n \overline{f_n}} = p_n \pm \sqrt{(p_n - \overline{f_n})^2 + \sigma_n^2}. \qquad (2.21b)$$

The interesting aspect of this result is that the choice of the typifying vector E depends on P, the vector to be compared to the set. This fact does not render E any less significant. Instead of comparing P with every member of the set, as in Equation 2.1, it is equivalent to compare it with E, given in Equation 2.21. The set of known members of the category appears in E as the constants $\overline{f_n^2}$ and $\overline{f_n}$, which may be computed once and for all. This fact has important implications regarding the amount of information which must be stored. In the comparison of an arbitrary point P with the set $\{F_n\}$ by means of the function $S(P, \{F_n\})$, all M members of the set must be stored, each having N coordinates. The total stored information about the set is thus MN numbers. In the comparison of P with E, on the other hand, the total storage is only 2N numbers.

2.4 Choosing the Optimum Orthogonal Coordinate System

The labeled events which belong to one category have been assumed given as vectors in an a priori selected coordinate system which expressed features of the events thought relevant to the determination of the category. An optimum set of feature weighting coefficients were then found through which similar events could be judged most similar to one another. It would be purely coincidental, however, if the features represented by the given coordinate system were optimal in expressing

the similarities among members of the set. In this section, therefore, we look for a new set of coordinates, spanning the same space, and expressing a different set of features which minimize the mean-square distance between members of the set. The problem just stated can be thought of as either enlarging the class of metrics considered thus far in the measure of similarity defined earlier or as enlarging the class of transformations of the space within which class we look for that particular transformation which minimizes the mean-square distance between similar events.

It was proved earlier that the linear transformation which changes the scale of the n^{th} dimension of the space by the factor w_{nn} while keeping the volume of the space constant and minimizing the mean-square distance between the transformed vectors is given by Equation 2.22.

$$F' = F[W], \quad \text{where } [W] = \begin{bmatrix} w_{11} & & & 0 \\ & w_{22} & & \\ & & \ddots & \\ 0 & & & w_{NN} \end{bmatrix} \tag{2.22a}$$

and

$$w_{nn} = \left(\prod_{p=1}^{N} \sigma_p \right)^{1/N} \frac{1}{\sigma_n} . \tag{2.22b}$$

The mean-square distance under this transformation is given by Equation 2.23 and is a minimum for the <u>given</u> choice of orthogonal coordinate system.

$$\overline{D^2} = \frac{1}{M(M-1)} \sum_{p=1}^{M} \sum_{m=1}^{M} \sum_{n=1}^{N} w_{nn}^{\,2} \, (f_{mn} - f_{pn})^2 = \text{minimum.} \tag{2.23}$$

It is possible, however, to rotate the coordinate system until one is found which minimizes the above minimum mean-square distance. Whereas the first minimization took place with respect to all choices of the w_{nn}'s, we are now interested in further minimizing this by first rotating the coordinate system so that the above optimum choice of w_{nn}'s should result in the absolute minimum distance between vectors. The solution of the above search for the optimum transformation may be conveniently stated in the form of the following theorem.

Theorem

The linear transformation which, after transformation, minimizes the mean-square distance between a set of vectors, subject to the constraint that the volume of the space is invariant under transformation, is a rotation $[C]$ followed by a diagonal transformation $[W]$. The rows of the matrix $[C]$ are eigenvectors of the covariance matrix $[U]$ of the set of vectors, and the elements of $[W]$ are those given in Equation 2.22b, where σ_p is the standard deviation of the coefficients of the set of vectors in the direction of the p^{th} eigenvector of $[U]$.

The proof of the above theorem is readily obtained as follows.

Proof

Expanding the square of Equation 2.23 and substituting the values of w_{nn} results in Equation 2.24 which is to be minimized over all choices of the coordinate system.

$$\overline{D^2} = \frac{1}{M(M-1)} \sum_{n=1}^{N} w_{nn}^2 \sum_{p=1}^{N} \sum_{m=1}^{M} (f_{mn}^2 + f_{pn}^2 - 2 f_{mn} f_{pn}) \tag{2.24a}$$

$$= \frac{2M}{(M-1)} \sum_{n=1}^{N} w_{nn}^2 \left(\overline{f_n^2} - \overline{f_n}^2\right) = \frac{2M}{(M-1)} \sum_{n=1}^{N} w_{nn}^2 \sigma_n^2 \tag{2.24b}$$

$$= \frac{2M}{(M-1)} \sum_{n=1}^{N} w_{nn}^2 \sigma_n^2 = \frac{M}{(M-1)} 2N \left[\prod_{p=1}^{N} \sigma_p^2\right]^{1/N} \tag{2.24c}$$

Let the given coordinate system be transformed by the matrix $[C]$.

$$[C] = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1N} \\ c_{21} & c_{22} & \cdots & c_{2N} \\ \cdot & \cdot & \cdot & \cdot \\ c_{N1} & c_{N2} & \cdots & c_{NN} \end{bmatrix}, \text{ where } \sum_{n=1}^{N} c_{pn}^2 = 1 \quad \text{for } p=1, 2, \ldots, N. \tag{2.25}$$

Equation 2.24 is minimized if the bracketed expression in Equation 2.24c is minimized. The latter may be named β and written as below.

$$\beta = \prod_{p=1}^{N} \sigma_p^2 = \prod_{p=1}^{N} \left[\frac{1}{M} \sum_{m=1}^{M} (f'_{mp})^2 - \left(\frac{1}{M} \sum_{m=1}^{M} f'_{mp}\right)^2\right], \tag{2.26a}$$

where

$$f'_{mp} = \sum_{n=1}^{N} f_{mn} c_{pn}. \tag{2.26b}$$

Substituting Equation 2.26b into 2.26a, we obtain Equation 2.27, where the averaging is understood to be over the set of M vectors.

$$\beta = \prod_{p=1}^{N} \left[\sum_{n=1}^{N} \sum_{s=1}^{N} \frac{1}{M} \sum_{m=1}^{M} f_{mn} f_{ms} c_{pn} c_{ps} - \left(\sum_{n=1}^{N} \overline{f}_n c_{pn}\right)^2\right]. \tag{2.27}$$

The squared expression may be written as a double sum and the entire equation simplified to 2.28.

$$\beta = \prod_{p=1}^{N} \sum_{n=1}^{N} \sum_{s=1}^{N} (\overline{f_n f_s} - \bar{f}_n \bar{f}_s)\, c_{pn}\, c_{ps}. \tag{2.28}$$

But $(\overline{f_n f_s} - \bar{f}_n \bar{f}_s) = u_{ns} = u_{sn}$ is an element of the covariance matrix $[U]$. Hence

$$\beta = \prod_{p=1}^{N} \sum_{n=1}^{N} \sum_{s=1}^{N} u_{ns} c_{pn} c_{ps}. \tag{2.29}$$

Using the method of Lagrange multipliers to minimize β in Equation 2.29, subject to the constraint of Equation 2.25, we obtain Equation 2.30 below as the total differential of β. The differential of the constraint, γ, is given in Equation 2.31.

$$d\beta(c_{11} c_{12} \ldots c_{NN}) = \sum_{\ell=1}^{N} \sum_{g=1}^{N} \left[\prod_{p \neq \ell}^{N} \sum_{n=1}^{N} \sum_{s=1}^{N} u_{ns} c_{pn} c_{ps} \right] \frac{\partial}{\partial c_{\ell g}} \left(\sum_{a=1}^{N} \sum_{b=1}^{N} u_{ab} c_{\ell a} c_{\ell b} \right) dc_{\ell g} = 0. \tag{2.30}$$

$$d\gamma = 2 \sum_{g=1}^{N} c_{\ell g}\, dc_{\ell g} = 0, \quad \text{for } \ell = 1, 2, \ldots, N. \tag{2.31}$$

In the way of an explanation of Equation 2.30, it is seen that when Equation 2.29 is differentiated with respect to $c_{\ell q}$, then all the factors in the product in Equation 2.29, where $p \neq \ell$, are simply constants. Carrying out the differentiation stated in Equation 2.30, we obtain

$$d\beta = \sum_{\ell=1}^{N} \sum_{g=1}^{N} dc_{\ell g} \left[\sum_{b=1}^{N} c_{\ell b} u_{gb} \right] \prod_{p \neq \ell}^{N} \left[\sum_{n=1}^{N} \sum_{s=1}^{N} u_{ns} c_{pn} c_{ps} \right] = 0. \qquad (2.32)$$

$$\text{Let } \prod_{p \neq \ell}^{N} \sum_{n=1}^{N} \sum_{s=1}^{N} u_{ns} c_{pn} c_{ps} = A_{\ell}. \qquad (2.33)$$

Note that since $p \neq \ell$, A_{ℓ} is just a constant as regards optimization of any $c_{\ell x}$.

In accordance with the method of Lagrange multipliers, each of the N constraints of Equation 2.31 is multiplied by a different arbitrary constant B_{ℓ} and is added to $d\beta$ as shown below.

$$d\beta + \sum_{\ell=1}^{N} B_{\ell} d\gamma_{\ell} = 0 = \sum_{\ell=1}^{N} \sum_{g=1}^{N} dc_{\ell g} \left[\left(\sum_{b=1}^{N} c_{\ell b} u_{gb} \right) A_{\ell} + B_{\ell} c_{\ell g} \right] = 0. \qquad (2.34)$$

By letting $-\lambda_{\ell} = B_{\ell} / A_{\ell}$ and by recognizing that $dc_{\ell q}$ is arbitrary, we get

$$\sum_{b=1}^{N} c_{\ell b} u_{gb} - \lambda_{\ell} c_{\ell g} = 0, \quad \text{for } g=1, 2, \ldots, N \text{ and } \ell=1, 2, \ldots, N. \qquad (2.35)$$

Let the ℓ^{th} row of the $[C]$ matrix be the vector C_{ℓ}. Then the above equation may be written as the eigenvalue problem of Equation 2.36 by recalling that $u_{qb} = u_{bq}$.

$$C_{\ell} \left[U - \lambda_{\ell} I \right] = 0, \quad \text{for } \ell = 1, 2, \ldots, N. \qquad (2.36)$$

Solutions of Equation 2.36 exist only for N specific values of λ_ℓ. The vector C_ℓ is an eigenvector of the covariance matrix $[U]$. The eigenvalues λ_ℓ are positive and the corresponding eigenvectors are orthogonal since the matrix $[U]$ is positive definite. Since the transformation $[C]$ is to be non-singular, the different rows C_ℓ must correspond to different eigenvalues of $[U]$. It may be shown that the only extremum of β is a minimum, subject to the constraint of Equation 2.25. Thus the optimum linear transformation which minimizes the mean-square distance of a set of vectors while keeping the volume of the space constant is given by Equation 2.37, where rows of $[C]$ are eigenvectors of the covariance matrix $[U]$.

$$\begin{bmatrix} a_{11}a_{12}\cdots a_{1N} \\ a_{21}a_{22}\cdots a_{2N} \\ \cdots\cdots\cdots \\ a_{N1}a_{N2}\cdots a_{NN} \end{bmatrix} = \begin{bmatrix} c_{11}c_{12}\cdots c_{1N} \\ c_{21}c_{22}\cdots c_{2N} \\ \cdots\cdots\cdots \\ c_{N1}c_{N2}\cdots c_{NN} \end{bmatrix}^T \begin{bmatrix} w_{11} & & & \\ & w_{22} & & \\ & & \ddots & \\ & & & w_{NN} \end{bmatrix} = \begin{bmatrix} w_{11}c_{11} & w_{22}c_{21} & \cdots & w_{NN}c_{N1} \\ w_{11}c_{12} & w_{22}c_{22} & \cdots & w_{NN}c_{N2} \\ \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots \\ w_{11}c_{1N} & w_{22}c_{2N} & \cdots & w_{NN}c_{NN} \end{bmatrix} \quad (2.37)$$

The numerical value of the minimum mean-square distance may now be computed as follows. The quantity $\overline{D^2}$ was given in Equation 2.24c which is reproduced here as Equation 2.38.

$$\overline{D^2} = \frac{M}{(M-1)}\, 2N \left[\prod_{p=1}^{N} \sigma_p^{\,2} \right]^{1/N} = \frac{M}{(M-1)}\, 2N \cdot (\beta)^{1/N}. \quad (2.38)$$

Substituting β from Equation 2.29, we obtain Equation 2.39.

$$\overline{D^2} = \frac{M}{(M-1)} 2N \left[\prod_{p=1}^{N} \sum_{n=1}^{N} \sum_{s=1}^{N} u_{ns} c_{pn} c_{ps} \right]^{1/N}. \tag{2.39}$$

But from Equation 2.35 we see that min $\overline{D^2}$ may be written as below, where the constraint 2.25 has also been utilized.

$$\min \overline{D^2} = \frac{M}{(M-1)} 2N \left[\prod_{p=1}^{N} \sum_{n=1}^{N} \lambda_p \, c_{pn}^{\,2} \right]^{1/N} = \frac{M}{(M-1)} 2N \left(\prod_{p=1}^{N} \lambda_p \right)^{1/N}. \tag{2.40}$$

It should be noted that the constraint of Equation 2.25 is not, in general, a constant volume constraint. It is that only if the transformation $[C]$ is orthogonal, as is the case in the solution just obtained. The set of transformations which keeps the volume constant is T_v in Figure 5. A subset of these are the orthogonal transformations T_o of constant volume, of which the optimum was desired. The solution presented here found the optimum transformation among a set of T_L which contains orthogonal transformations of constant volume but is not necessarily constant volume for those which are non-orthogonal. The solution here given, therefore, is optimum among the constant volume transformations $T_v \cap T_L$ shown shaded in Figure 5. This intersection is a larger set of transformations than that for which the optimum was sought.

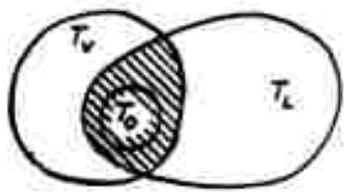


Figure 5. Sets of Transformations

The methods of this chapter are optimal in measuring membership in categories of certain types. Suppose, for instance, that categories are statistically independent random processes which generate members with multivariate Gaussian probability distributions of unknown means and variances. Elsewhere it is shown that the metric developed here measures contours of equal a posteriori probabilities. Given the set of labeled events, the metric specifies the locus of points which are members of the category in question with equal probability.

Before bringing this chapter to a conclusion, the important concepts introduced here will be summarized.

Categorization, the basic problem of pattern recognition, is regarded as the process of learning how to partition the signal space into regions where each contains points of only one category. The notion of similarity between a point and a set of points of a category plays a dominant role in the partitioning of signal space. Similarity of a point to a set of points is regarded as the average "distance" between the point and the set. The sense in which distance is understood is not specified, but the optimum sense is thought to be that which (by the optimum method of measuring distance) clusters most highly those points which belong to the same category. The mean-square distance between points of a category is a measure of clustering. An equivalent alternate interpretation of similarity (not as general as the interpretation above) is that the transformation which optimally clusters like points, subject to suitable criteria to assure the non-triviality of the transformations, is instrumental in exhibiting the similarities between points of a set. In particular, the optimum orthogonal transformation and hence a non-Euclidean method of measuring distance is found which minimizes the mean-square distance between a set of points, if the volume of the space is held constant to assure

non-triviality. The resulting measure of similarity between a point P and a set $\{F_m\}$ is given in Equation 2.41, where a_{ns} is given the Theorem of this chapter.

$$S(P, \{F_m\}) = \frac{1}{M} \sum_{m=1}^{M} \sum_{n=1}^{N} \left[\sum_{s=1}^{N} a_{ns}(p_s - f_{ms}) \right]^2 . \qquad (2.41)$$

To facilitate the instrumentation of computations of the function S, a typifying example E of the set is developed which sets an upper bound on the necessary information storage at 2N, numbers, where N is the number of dimensions of the space in which the points are represented.

3. CATEGORIZATION

3.1 The Process of Classification

Pattern recognition consists of the twofold task of "learning", on one hand, what the category or class is to which a set of events belongs, and of deciding, on the other hand, whether a new event belongs to the category or not. In this chapter details of the method of accomplishing these two parts of the task are discussed, subject to the limitations on recognizable categories imposed by the assumptions stated earlier.

In the following section two distinct modes of operation of the recognition system will be distinguished. The first of these consists of the sequential introduction of a set of events, each labeled according to the category to which it belongs. During this period, identification of the common pattern of the inputs which allow their classification into their respective categories is desired. As part of the process of learning to categorize, the estimate of what the category is must also be updated to include each new event as it is introduced. The process of updating the estimate of the common pattern consists of recomputing the new measures of similarity and the typifying examples of the sets so that these will include the new, labeled event on which the above quantities are based.

During the second mode of operation the event P to be classified is compared to each of the sets of labeled events by the measure of similarity found best for each set. The event is then classified as a member of that category to which it is most similar.

It is not possible to state with certainty that the pattern has been successfully learned or recognized from a set of its examples, because

information is not available on how examples were selected to represent the class. Nevertheless, it is possible to obtain a qualitative indication of how certain we may be of having obtained a correct method of determining membership in the category from the ensemble of similar events. As each new event is introduced, its similarity to the members of the sets already presented is measured by the function S defined in the preceding chapter. The magnitude of the number S indicates how close the new event is to those already introduced. As S is refined and, with each new example improves its ability to recognize the class, the numerical measure of similarity between new examples and the class will tend to decrease, on the average. Strictly speaking, of course, this last statement cannot be true in general. It may be true only if the categories to be distinguished are separable by functions S taken from the class which we have considered; even under this condition the statement is true only if certain assumptions are made regarding the statistical distribution of the samples on which we learn. Since we have no a priori knowledge regarding the satisfaction of either of these two requirements, the convergence of the similarity as the sample size is increased is simply qualitative wishful thinking whose heuristic justification is based on the minimization problem solved in developing S.

Figure 6 illustrates the mechanization of the learning and recognition modes of the special classificatory process discussed so far. For the sake of clarity, the elementary block diagram of the process is shown to distinguish only between two categories of events, but it can be extended readily to distinguish between an arbitrary number of categories. It should be noted that one of the categories may be the complement of all others.

The admission of such a category into the set is one of the ways in which a machine which is always forced to classify events into known categories may be made to decide that an event does not belong to any of the desired ones; it belongs to the category of "everything else". Samples of "everything else" must, of course, be given.

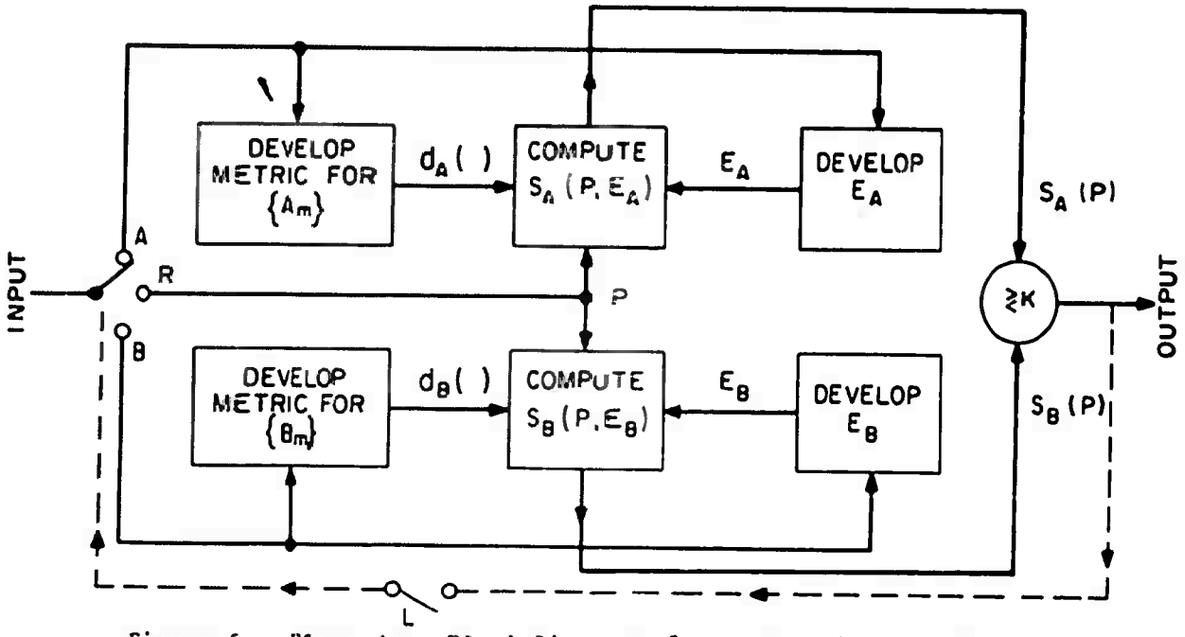


Figure 6. Elementary Block Diagram of the Classification Process

During the first mode of operation, the input to the machine is a set of labeled events. Let us follow its behavior through an example. Suppose that a number of events, some belonging to set A and some to set B, have already been introduced. According to the method described in the previous chapter, therefore, the optimum metrics (one for each class) have been found which minimize the mean-square distance between events of the same set. Similarly, the best exemplars of the sets have also been found. As a new labeled event is introduced (say, it belongs to set A), the switch at the input is first turned to the recognition mode R so that the new event P may be

compared to set A as well as to set B through the functions $S_A(P) = S(P, \{A_m\}) = S_A(P, E_A)$ and $S_B(P)$ which were computed before the introduction of P. The comparison of $S_A - S_B$ with a threshold K indicates whether the point P would be classified correctly or incorrectly from knowledge available up to the present. The input switch is then turned to A so that P, which indeed belongs to A, may be included in the computation of the best metric and exemplar of set A.

When the next labeled event is introduced (let us say it belongs to set B), the input switch is again turned to B to test the ability of the machine to classify the new event correctly. After the test, the switch is turned to B so that the event may be included among the examples of set B and the optimum function S_B may be recomputed. This procedure is repeated for each new event, and a record is kept of the rate at which incorrect classifications would be made on the known events. When the training period is completed, presumably as a result of satisfactory performance on the selection of known events, the input switch is left in the recognition mode.

3.2 Learning

"Supervised learning" takes place in the interval of time in which examples of the categories generate ensembles of points from which the defining features of the classes are obtained by methods previously discussed. "Supervision" is provided by an outside source such as a human who elects to teach the recognition of pattern by examples, and who selects the examples on which to learn.

"Unsupervised learning", by contrast, is a method of learning without the aid of such an outside source. It is clear, at least intuitively, that

the unsupervised learning of membership in specific classes cannot succeed unless it is preceded by a period of supervision, during which some concepts regarding the characteristics of classes are established. A specified degree of certainty concerning the patterns has been achieved in the form of a sufficiently low rate of misclassification during the supervised learning period. The achievement of the low misclassification rate, in fact, can be used to signify the end of the learning period, after which the system which performs the operations indicated in Figure 6 may be left to its own devices. It is only after this supervised interval of time that the system may be usefully employed to recognize, without outside aid, events as belonging to one or another of the categories.

Throughout the period of learning on examples, each example is included in its proper set of similar events which influence the changes of the measures of similarity. After supervised activity has ceased, events introduced for classification may belong to any of the categories; and no outside source informs the machine of the correct category. The machine itself, operating on each new event, however, can determine, with the already qualitatively specified probability of error, to which class the event should belong. If the new event is included in the set exemplifying this class, the function measuring membership in the category has been altered. Unsupervised learning results from the successive alterations of the metrics, brought about by the inclusion of events into the sets of labeled events according to determination of class membership rendered by the machine itself. This learning process is instrumented by the dotted line in Figure 6 which,

when the learning switch L is closed, allows the machine's decisions to control routing of the input to the various sets.

To facilitate the illustration of some implications of the process described above, consider the case in which recognition of membership in a single class is desired and all the labeled events are members of only that class. In this case, classification of events as members or nonmembers of the category degenerates into the comparison of the similarity S with a threshold T. If S is greater than T, the event is a nonmember; if S is less than T, on the other hand, the event is said to be a member of the class. Since the machine decides that all points of the signal space for which S is less than T are members of the class, the latter, as far as the machine is concerned, is the collection of points which lie in a given region in the signal space. For the specific function S of the previous chapter, this region is an ellipsoid in the N-dimensional space.

Unsupervised learning is graphically illustrated in Figure 7. The two-dimensional ellipse drawn with a solid line signifies the domain D_1 of the signal space in which any point yields $S < T$. This domain was obtained during supervised activity. If a point P_1 is introduced after supervised learning, so that P_1 lies outside D_1, then P_1 is merely rejected as a nonmember of the class. If point P_2 contained in D_1 is introduced, however, it is judged a member of the class and is included in the set of examples to generate a new function S and a new domain D_2, designated by the dotted line in Figure 7. A third point P_3 which was a nonmember before the introduction of P_2 becomes recognized as member of the class after the inclusion of P_2 in the set of similar events.

Figure 7. Unsupervised Learning

Although the tendency of this process of "learning" is to perpetuate the original domain, it has interesting properties worth investigating. The investigation of unsupervised learning would form the basis for a valuable continuation of the work presented herein.

Before leaving the subject of unsupervised learning, it should be pointed out that as the new domain D_2 is formed, points such as P_4 in Figure 7 become excluded from the class. Such an exclusion from the class is analogous to "forgetting" because of lack of repetition. Forgetting is the characteristic of not recognizing P_4 as a member of the class, whereas at one time it was recognized to belong to it.

3.3 Threshold Setting

In the classification of an event P the mean-square distance between P and members of each of the categories is computed. The distance between P and members of a category C is what we called "similarity", $S_C(P)$, where the "sense" in which "distance" is understood depends on the particular category in question. We then stated that, in a manner analogous to decisions

based on maximum likelihood ratios, the point P is classified as a member of the category to which it is most similar. Hence, P belongs to category C if $S_C(P) \leq S_X(P)$, where X is any of the other categories.

Since in this special theory the function $S_C(P)$ which measures membership in category C, was developed by maximally clustering points of C without separating them from points of other sets, there is no guarantee, in general, that a point of another set B may not be closer to C than to B. This is guaranteed only if points of the sets satisfy certain conditions which will be stated below. A graphical illustration which clarifies the comparison of similarities of a point to the different categories is shown in Figure 8. In this figure the elliptical contours $S_{A_1}(P)$, $S_{A_2}(P)$, etc., indicate the loci of points P in the signal space which are at a mean-square distance of 1, 2,....,etc., from members of category A. The loci of these points are concentric ellipsoids in the N-dimensional signal space, shown here in only two dimensions. Similarly, $S_{B_1}(P)$, $S_{B_2}(P)$,..... etc., and $S_{C_1}(P)$, $S_{C_2}(P)$,...., etc., are the loci of those points whose mean-square distance from categories B and C, respectively, are 1, 2,...., etc. Note carefully that the sense in which distance is measured to each of the categories differs as is indicated by the different orientations and eccentricities of the ellipses.

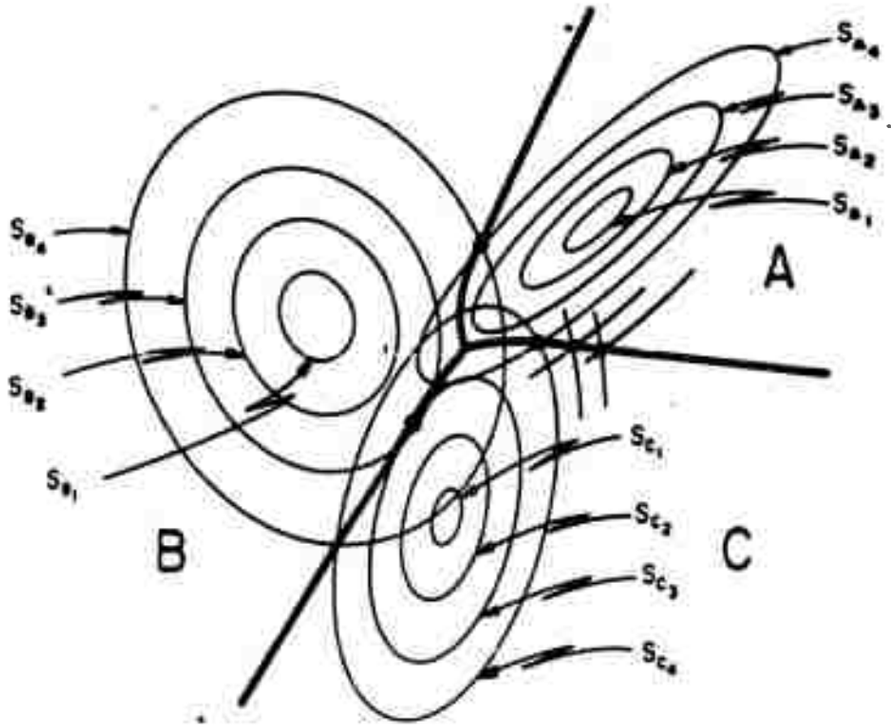


Figure 8. Categorization

The heavy line shows the loci of points which are at equal mean-square distances to two or more sets according to the manner in which distance is measured to each set. This line, therefore, defines the boundary of each of the categories.

At this point in the discussion it would be helpful to digress from the subject of thresholds and dispel some misconceptions which Figure 8 might create regarding the general nature of the categories found with the method described herein. It will be recalled that one of the possible

ways in which a point not belonging to either category could be so classified was by allowing a separate category for "everything else" and assigning the point to the category to which its mean-square distance is smallest. Another, perhaps more practical, method is to call a point a member of neither category if its mean-square distance to the set of points of any class exceeds some threshold value. If this threshold value is set, for example, at a mean-square distance of 3 for all of the categories in Figure 8, then points belonging to A, B, and C will lie inside the three ellipses shown in Figure 9.

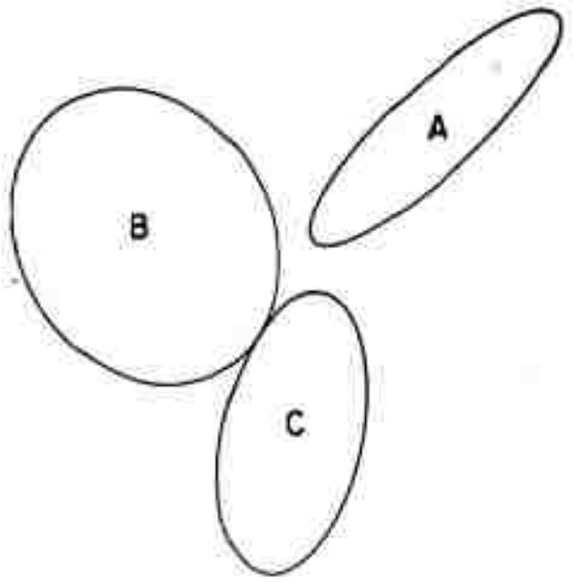


Figure 9. Categorization with Threshold

It is readily seen, of course, that there is no particular reason why one given minimum mean-square distance should be selected instead of another; or, for that matter, that this minimum distance be the same for all categories. Many logical and useful criteria may be selected for determining the optimum threshold setting. Here, only one criterion will be singled out as

particularly useful. This criterion requires that the minimum thresholds be set so that most of the labeled points fall into the correct category. This is a fundamental criterion, for it requires the system to be designed to work best by making the largest number of correct decisions.

The criterion of selecting a threshold to make the most correct classifications may be applied to our earlier discussion where the boundary between categories was determined by equating the similarities of a point to two or more categories. In the particular example of Figure 6, where a point could be a member of only one of two categories A and B, the difference $S_A - S_B = 0$ formed the dividing line. There is nothing magical about the threshold zero; one might require that the dividing line between the two categories be $S_A - S_B = K$, where K is a constant chosen from other considerations. A similar problem in communication theory is the choice of a signal-to-noise ratio which serves as the dividing line between calling the received waveform "signal" or calling it "noise". It is understood, of course, that signal-to-noise ratio is an appropriate criterion on which to base decisions (at least in some cases), but the particular value of the ratio to be used as a threshold level must be determined from additional requirements. In communication theory these are usually requirements on the false alarm or false dismissal rates. In the problem of choosing the constant K, we may require that it be selected so that most of the labeled points lie in the correct category.

3.4 Practical Considerations

In considering the instrumentation of the process of categorization

previously described, two main objectives of the machine design must receive careful consideration. The first of these is the practical requirement that all computations involved in either the learning or the recognition mode of the machine's operation be performed as rapidly as possible. It is especially desirable that the classification or recognition of a new event be implemented in essentially real time. The importance of this requirement is readily appreciated if the classificatory technique is considered in terms of an application such as the automatic recognition of speech events, an important part of voice controlled phonetic typewriters. The second major objective, not unrelated to the first, is that the storage capacity required of the machine have an upper bound, thus assuring that the machine is of finite and predetermined size. At first glance it seems that the instrumentation of the machine of Figure 6 requires a storage capacity proportional to the number of events encountered during the machine's experience. This seems so because the set of labeled events on which the computations are carried out must be stored in the machine. It will be shown in this section, however, that all computations may be performed from knowledge of only certain statistics of the set of labeled events, and that these statistics may be recomputed to include a new event without knowledge of the original set. Therefore, it is necessary to store only these statistics, the number of which is independent of the number of points in the set.

It will be recalled that there are two instances where knowledge of the data matrix is necessary. The data matrix $[F]$, given in Equation 3.1, is the M x N matrix of coefficients which results when the M given examples of the same category are represented as N-dimensional vectors.

$$[F] = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1N} \\ f_{21} & f_{22} & \cdots & f_{2N} \\ \cdot & \cdot & \cdots & \cdot \\ f_{M1} & f_{M2} & \cdots & f_{MN} \end{bmatrix} \qquad (3.1)$$

The first use of this matrix occurs in the computation of the optimum orthogonal transformation or metric which minimizes the mean-square distance of the set of like events. This transformation is stated in the Theorem in Section 2.4 and is given in Equation 2.37 as the product of an orthonormal and diagonal transformation. Rows of the orthonormal transformation $[C]$ are eigenvectors of the covariance matrix $[U]$ computed from the data matrix of Equation 3.1, and elements of the diagonal matrix $[W]$ are the reciprocal standard deviations of the data matrix after it has been transformed by the orthonormal transformation $[C]$.

The second use of the matrix $[F]$ occurs when an unclassified event P is compared to the set by measuring the mean-square distance between P and points of the set after both the point and the set have been transformed. This latter comparison is replaced by the measurement of the distance between the transformed point P and a "typical example" of the set, as stated by Equation 2.19. The quantities of interest in this computation, as seen from Equation 2.21, are the mean, the mean-square, and the standard deviation of the elements in the columns of the data matrix <u>after</u> the orthonormal transformation.

Reduction of the necessary storage facility of the machine may be accomplished if only the covariance matrix, the means, the mean-squares, and the standard deviations of the transformed data matrix are used in the computations, and if these may be recomputed without reference to the original data matrix. The expression of the above quantities when based on M+1 events may be computed from the corresponding quantity based on M events and a complete knowledge of the M+1st event itself. The method of the computations is described below.

(1) The covariance matrix of M+1 events.

The general coefficient of the covariance matrix $[U]$ of the set of events given by the data matrix $[F]$ is given in Equation 3.2

$$u_{ns} = u_{sn} = \overline{f_n f_s} - \overline{f}_n \overline{f}_s \tag{3.2}$$

Note, incidentally, that the matrix $[U]$ may be written as in Equation 3.3, where the matrix $[J]$ has been introduced for convenience. As a check, let us compute the general element u_{ns}

$$[U] = \frac{1}{M} [F - J]^T [F - J] \tag{3.3a}$$

$$\text{where } [J] = \begin{bmatrix} \overline{f}_1 & \overline{f}_2 & \cdots & \overline{f}_N \\ \overline{f}_1 & \overline{f}_2 & \cdots & \overline{f}_N \\ \cdot & \cdot & \cdots & \cdot \\ \overline{f}_1 & \overline{f}_2 & \cdots & \overline{f}_n \end{bmatrix} \tag{3.3b}$$

The n^{th} column of the $[F - J]$ matrix, which becomes the n^{th} row of its transpose, is given in Equation 3.4 as well as the s^{th} column of $[F - J]$.

The product is the covariance matrix coefficient u_{ns}.

$$u_{ns} = \frac{1}{N}\left[\left(f_{1n}-\bar{f}_1\right)\cdot\left(f_{2n}-\bar{f}_2\right)\cdots\left(f_{Mn}-\bar{f}_M\right)\right]\begin{bmatrix} f_{1s}-\bar{f}_s \\ f_{2s}-\bar{f}_s \\ \vdots \\ f_{Ms}-\bar{f}_s \end{bmatrix} \tag{3.4}$$

$$u_{ns} = \frac{1}{N}\sum_{m=1}^{M}\left(f_{mn}-\bar{f}_n\right)\left(f_{ms}-\bar{f}_s\right) = \overline{f_n f_s} - \bar{f}_n\bar{f}_s \tag{3.5}$$

Now to compute the covariance based on M+1 events, $u_{ns}(M+1)$, it is convenient to store the N means $\bar{f}_n$ for all values of n. It is also convenient to store the N(N+1)/2 independent values of $\overline{f_n f_s}$. Both of these quantities may be updated readily as a new event is introduced. The mean $\bar{f}_n^{M+1}$ based on M+1 events may be obtained from the mean based on only M events, f_n, from Equation 3.6a and $\overline{f_n f_s}^{M+1}$ may be obtained from Equation 3.6b.

$$\bar{f}_n^{M+1} = \frac{M\,\bar{f}_n^{M} + f_{M+1,n}}{M+1} \tag{3.6a}$$

$$\overline{f_n f_s}^{M+1} = \frac{M\,\overline{f_n f_s}^{M} + f_{M+1,n}\,f_{M+1,s}}{M+1} \tag{3.6b}$$

Here, the superscript of the ensemble average indicates the number of events partaking in the averaging, and $f_{M+1,n}$ is the n^{th} coefficient of the M+1st event. We now have everything necessary to compute the new covariance coefficients. The storage facility required thus far is N(N+3)/2+1 locations. The +1 is used for storing the number M. If the covariance matrix is also stored, the necessary number of storage locations is $(N+1)^2$; this makes use of the fact that both $[U]$ and $[F^T F]$ are symmetric matrices.

From the matrix $[U]$ the orthonormal transformation $[C]$ may be found by solving the eigenvalue problem $[C][u - \lambda I] = 0$. The matrix $[C]$ has to be stored, requiring an additional N^2 storage locations.

(2) Mean of the p^{th} column of the $[F][C] = [F']$.

As stated earlier, one of the quantities of interest in the typifying example is the mean of the elements in a column of the data matrix after its orthonormal transformation with $[C]$. The general element of the $[F']$ matrix is f'_{mp} given in Equation 2.26b and in 3.7a, and its mean is given in Equation 3.7b).

$$f'_{mp} = \sum_{n=1}^{N} f_{mn} c_{pn} \tag{3.7a}$$

$$\overline{f'_p} = \frac{1}{M} \sum_{m=1}^{M} \sum_{n=1}^{N} f_{mn} c_{pn} = \sum_{n=1}^{N} \overline{f}_n c_{pn} \tag{3.7b}$$

No additional storage is required to compute $\overline{f'_p}$, since all the factors of Equation 3.7b are already known. An additional N locations must be made available to store the N means, however.

(3) Mean-square of p^{th} column of $[F']$.

The mean-square value of elements of the p^{th} column of $[F']$ is given in Equation 3.8a and b.

$$\overline{f'^2_p} = \frac{1}{M} \sum_{m=1}^{M} f'^2_{mp} = \frac{1}{M} \sum_{m=1}^{M} \sum_{n=1}^{N} \sum_{s=1}^{N} f_{mn} f_{ms} c_{pn} c_{ps} \tag{3.8a}$$

$$f'^2_p = \sum_{n=1}^{N} \sum_{s=1}^{N} \overline{f_n f_s} c_{pn} c_{ps} \tag{3.8b}$$

No additional storage is necessary for this computation. An additional N locations, however, must be available to store $\overline{f'^2_p}$.

(4) Standard deviation of the p^{th} column of $[F']$.

The only remaining quantity necessary in the instrumentation of the recognition system is the reciprocal standard deviation of the p^{th} column of $[F']$, as stated in the theorem of Chapter 2. The standard deviation and the elements of the diagonal matrix $[W]$ are given by Equation 3.9, where all the quantities are already known. An additional N locations are needed to store their values, however.

$$W_{pp} \propto \frac{1}{\sigma'_p} = \frac{1}{\sqrt{\overline{f'^2_p} - \overline{f'_p}^2}} \tag{3.9}$$

The total number of storage locations is $2N^2 + 5N+1$ for each of the categories to which events may belong. If the number of examples M of a category is less than the number of dimensions N of the space in which they are represented, the required number of storage locations is only $2M^2 + 5M+1$. In order to utilize this further reduction of storage and computational time, however, the M events must be reexpressed in a new coordinate system obtained through the Schmidt orthogonalization of the set of M vectors representing the examples of the set. In the beginning of the learning process, when the number of labeled events is very much smaller than the number of dimensions of the space, the saving achieved by Schmidt orthogonalization is very significant.

A practical remark worthy of mention is that at the beginning of the learning process, when M is less than N, the solution of the eigenvalue problem $[U - \lambda I] = 0$ may be greatly simplified by recognition of the fact that $[U]$ is singular if M<N. Although it is not immediately obvious, nevertheless, it is true that the non-zero eigenvalues of $[U]$ in Equation 3.3a are identical to the eigenvalues of the matrix $(F - J)(F - J)^T$ as stated below.

Non-zero eigenvalues of $(F - J)^T(F-J)$ = eigenvalues of $(F - J)(F - J)^T$ (3.10)

The first of the matrices is an N x N, while the second is an M x M matrix. There are N-M zero eigenvalues of the first matrix; the computational advantage of working with the second matrix for M<N is therefore significant.

A few additional remarks should be made about the nature of the solution obtained with the two constraints of Equations 2.12b and 2.13 . It should be noted, first of all, that if the number of points in a set is equal to or less than the number of dimensions in which they are expressed, then a hyperplane of one less dimensions can always be passed through the points. Along any direction orthogonal to this hyperplane, the projections of points of the set F are equal. Along such a direction, therefore, the variance of the given points is zero, leading to a zero eigenvalue of the covariance matrix. This results in calling the corresponding eigenvector (the direction about which the variance is zero), an "all important" feature. The feature weighting coefficient W_n is thus unity or infinity, depending on which of the above two constraints were applied. If the second or constant volume constraint were used, each point of the set F used in learning would be correctly identified, and its distance to

the set F would be zero by the optimum metric. At the same time the metric classifies each point of another category G as a nonmember of F. A new member of category F, on the other hand, would probably be misclassified, since it is unlikely that the new member of F would have exactly the same projection along the eigenvector as the other members had displayed. This misclassification would not occur if the number of examples of the category F exceeded the number of dimensions in which they were expressed. There are several methods to prevent misclassification; for example, if the first constraint were applied, misclassification of members of F would not occur.

Another fact of some importance which should be brought to the reader's attention is the physical significance of the eigenvectors. The vector with the smallest eigenvalue or largest feature weighting coefficient designates that feature of members of the set in which the members are most similar. This is not equivalent to the feature which is most similar to members of the set. The former is a solution of a problem in which we wish to find a direction along which the projections of the set on the average, are most nearly the same. The second is a solution of a problem where we wish to find the direction along which the projections of the set are largest, on the average. The desired direction, in the first case, is the eigenvector of the covariance matrix with the smallest eigenvalue; in the second case, it is the eigenvector of the correlation matrix $\left[F^T F\right]$ with the largest eigenvalue. It can be shown that the latter problem is equivalent to finding the set of orthonormal functions in which a process is to be expanded so that the truncation error, which results when only a finite number of terms of the expansion are retained, should be minimized, on the average. The set

of functions having this property are eigenfunctions of the correlation function of the process, and they are arranged in the order of decreasing eigenvalues.

The important concepts of this chapter will now be summarized. Pattern recognition consists of the twofold task of "learning", on the one hand, what the category is to which a set of events belongs; and of deciding on the other hand, whether a new event belongs to the category or not. "Learning", for the simple situation where similarity to a class of things is determined solely from examples of the class, may be instrumented in the form of the diagram of Figure 6. In this diagram "learning" consists of the construction of metrics or the development of linear transformations which maximize the clustering of points which represent similar events. A distinction is made between "supervised learning" (learning on known examples of the class) and "unsupervised learning" (learning through use of the machine's own experience). In this connection it is stated that the convergence of a learning process to correct category recognition, in most cases, probably cannot be guaranteed. The problem of threshold setting for partitioning the signal space is likened to the similar problem in the detection of noisy signals, and may be solved as an extremum problem. Finally, some practical considerations of importance in the mechanization of the decision process are discussed. It is shown that only finite storage capacity is required of the machine which instruments the techniques, and that the amount of storage has an upper bound which depends on the number of dimensions of the signal space.

4. DECISION THEORETICAL BASIS OF CATEGORIZATION TECHNIQUES

The categorization techniques outlined in the preceding sections do not involve assumptions about the probability distributions of the respective categories. It is instructive, however, to consider the relation of these techniques to the conventional decision-theoretical approach to problems of categorization. In the latter it is assumed that probability density functions for each category are known. Given such functions, it is possible to set up optimum procedures for categorization. It will be shown that under certain conditions the criteria developed in this report are exactly those prescribed by decision theory when the distributions are known. The important fact that should be kept in mind is that the categorization techniques discussed in earlier sections do not require knowledge of the density functions. They provide procedures of categorization where there is no knowledge of such distributions.

The purpose of this section is to provide a corroboration of the techniques and to lay bare their relation to decision theory proper. That such a corroboration should occur so fortuitously after the development of the techniques is gratifying in that it gives support from a well-established mathematical theory.

In order to set down the relation between the categorization techniques of earlier chapters and decision theory, it will be necessary to state briefly some of the assumptions and results of the latter. These results will be stated without proof since their full exposition may be found in any text on decision theory.

For purposes of exposition, two categories will be treated. Generalization to K categories proceeds in a natural way, but tends to obscure the essential simplicity of the theory. Assume, then, that there are two categories to which it is desired to assign objects as yet uncategorized. The only direct knowledge available about a specific object is a set of n measurements made upon it. Furthermore, a probability density function for each cateogry is known such that, when integrated over a region A of the n-dimensional space spanned by the n measurements, it yields the probability that an object from a given cateogry will produce measurements falling in region A. That is, the probability that an object from category C_i is accompanied by n measurements that fall in A is given by

$$\int_A p_i(x)\, dx$$

where $p_i(x)$ is the probability density function for cateogry C_i and x represents the vector $(x_1, x_2, \ldots, x_n)$.

Let it also be assumed that a priori probabilities π_1 and π_2, are known which give the probability of occurrence of an object from C_1 and C_2, respectively. The decision theory approach involves dividing the n-dimensional space into two regions, R_1 and R_2, such that when a set of measurements falls in R_1 the object is assigned to C_1 and, similarly, when the measurements fall in R_2, the object is assigned to C_2.

If the a priori probabilities and the density functions are known, then these regions may be chosen in such a way that the expected cost of making decisions is minimized. Here, it is assumed that there is a cost connected with making a misclassification. A division of n-dimensional space

into two regions R_1 and R_2 is called a decision procedure. The expected cost may be written

$$E(K) = \pi_1 K_{21} \int_{R_2} p_1(x)\, dx + \pi_2 K_{12} \int_{R_1} p_2(x)\, dx, \qquad (4.1)$$

where K_{21} is the cost of misclassifying an object from C_1; and K_{12}, that of misclassifying an object from C_2. The first term of Equation (4.1) is the expected cost due to misclassifying objects from C_1. Since $p_1(x)$ is the density function for C_1, its integration over R_2 (the region where the procedure specifies that the object be assigned to C_2) gives this expected cost. A similar statement may be made for the second term of Equation (4.1). Hence (4.1) gives the total expected cost.

It is desired to choose the regions R_1 and R_2 that minimize Equation (4.1). To determine these regions, we rewrite Equation (4.1) in the following manner.

$$E(K) = \int_{R_2} \left[\pi_1 K_{21} p_1(x) - \pi_2 K_{12} p_2(x)\right] dx + \int \pi_2 K_{12} p_2(x)\, dx \qquad (4.2)$$

The last term of Equation (4.2) is a positive number. Consequently, Equation (4.2) is made smaller by choosing the region R_2 so that it contains all (and only) those points x such that

$$\pi_1 K_{21} p_1(x) - \pi_2 K_{12} p_2(x) < 0. \qquad (4.3)$$

Thus, R_1 must be the region of points which satisfy

$$\pi_1 K_{21} p_1(x) - \pi_2 K_{12} p_2(x) \geq 0. \qquad (4.4)$$

Another way of writing Equations (4.3) and (4.4) is

$$R_1: \quad \frac{p_1(x)}{p_2(x)} \geq \frac{\pi_2 K_{12}}{\pi_1 K_{21}}, \tag{4.5}$$

$$R_2: \quad \frac{p_1(x)}{p_2(x)} < \frac{\pi_2 K_{12}}{\pi_1 K_{21}}. \tag{4.6}$$

The optimum dscision procedure when a priori probabilities, π_1 and π_2, and density functions are known is given by Equations (4.5) and (4.6). That is, given a sst of reasurements x the object represented by x is assigned to C_1 or C_2 depending on whether x satisfies inequality (4.5) or (4.6).

Unfortunatsly, the a priori probability of occurrencs of an object from a specified sst is seldom known. In lieu of these probabilities there are procedures which psrmit determination of the regions R_1 and R_2. Thus, ons might assume ths a priori probabilities to be equal. This is known as a maximum likelihood criterion. Another criterion is to minimize the maximum probability of misclassification. This is the "minimax" criterion. It is obtained by choosing the regions in such a manner that the expected cost of misclassifying an object from C_1 is equal to that of misclassifying an object from C_2, i.e.,

$$K_{12} \int_{R_1} p_2(x)\, dx = K_{21} \int_{R_2} p_1(x)\, dx.$$

We nsxt consider x distributed normally with mean μ_1 and covariance U_1 whsn it is a member of C_1, and msan μ_2 and covariance U_2 when

it is a member of C_2. That is, $p_1(x)$ is given by

$$p_1(x) = \frac{1}{(2\pi)^{1/2n}|U_1|^{1/2}} \exp -\tfrac{1}{2}(x-\mu_1)\,U_1^{-1}\,(x-\mu_1)' \qquad (4.7)$$

and $p_2(x)$, by

$$p_2(x) = \frac{1}{(2\pi)^{1/2n}|U_2|^{1/2}} \exp -\tfrac{1}{2}(x-\mu_2)\,U_2^{-1}\,(x-\mu_2)' \qquad (4.8)$$

The regions R_1 and R_2 as given by Equations (4.5) and (4.6) are

$$R_1: \frac{p_1(x)}{p_2(x)} = \frac{|U_2|^{1/2}}{|U_1|^{1/2}} \exp -\tfrac{1}{2}(x-\mu)\,U_1^{-1}\,(x-\mu_1)' + \tfrac{1}{2}(x-\mu_2)\,U_2^{-1}\,(x-\mu_2)' \geqslant \frac{\pi_2 K_{12}}{\pi_1 K_{21}}$$

$$R_1: \frac{p_1(x)}{p_2(x)} = \frac{|U_2|^{1/2}}{|U_1|^{1/2}} \exp -\tfrac{1}{2}(x-\mu_1)\,U_1^{-1}\,(x-\mu_1)' + \tfrac{1}{2}(x-\mu_2)\,U_2^{-1}\,(x-\mu_2)' < \frac{\pi_2 K_{12}}{\pi_1 K_{21}}$$

Since the logarithmic function is monotonically increasing, the ratio $\frac{p_1(x)}{p_2(x)}$ may be replaced by its logarithm, i.e.,

$$R_1: \log \frac{|U_2|^{1/2}}{|U_1|^{1/2}} - \tfrac{1}{2}\left[(x-\mu_1)\,U_1^{-1}\,(x-\mu_1)' - (x-\mu_2)\,U_2^{-1}\,(x-\mu_2)'\right] \geqslant \log \frac{\pi_2 K_{12}}{\pi_1 K_{21}} \qquad (4.9)$$

$$R_2: \log \frac{|U_2|^{1/2}}{|U_1|^{1/2}} - \tfrac{1}{2}\left[(x-\mu_1)\,U_1^{-1}\,(x-\mu_1)' - (x-\mu_2)\,U_2^{-1}\,(x-\mu_2)'\right] < \log \frac{\pi_2 K_{12}}{\pi_1 K_{21}} \qquad (4.10)$$

It will now be shown that the regions expressed in inequalities (4.9) and (4.10) are the same as those developed in the preceding sections for the categorization of unlabeled objects. First, let it be noted that

$$A_1 U_1 A_1' = \Lambda_1$$

where A_1 is the matrix the rows of which are the eigenvectors of U_1 and Λ_1 is the diagonal matrix of eigenvalues. Then

$$A_1 U_1^{-1} A_1' = \Lambda_1^{-1}$$

and

$$U_1^{-1} = A_1' \Lambda_1^{-1} A_1$$

Likewise, when A_2 and Λ_2 are the matrices of eigenvectors and eigenvalues of U_2,

$$U_2^{-1} = A_2' \Lambda_2^{-1} A_2 .$$

Hence,

$$-\frac{1}{2}\left[(x-\mu_1) U_1^{-1} (x-\mu_1)' - (x-\mu_2) U_2^{-1} (x-\mu_2)'\right]$$

$$= -\frac{1}{2}\left[(x-\mu_1) A_1' \Lambda^{-1} A_1 (x-\mu_1)' - (x-\mu_2) A_2' \Lambda_2^{-1} A_2 (x-\mu_2)'\right]$$

$$= -\frac{1}{2}\left[(x A_1' - \mu_1 A_1') \Lambda_1^{-1} (x A_1' - \mu_1 A_1')' - (x A_2' - \mu_2 A_2^1) \Lambda_2^{-1} (x A_2^1 - \mu_2 A_2^1)'\right].$$

It has been shown in the preceding sections that the transformation which minimizes the mean-square distance of the first category when volume is held invariant is given by

$$y = x A_1' \Lambda_1^{-1/2}, \qquad (4.12)$$

Similarly, the transformation that minimizes the mean-square distance of the second category is

$$y = x A_2' \Lambda_2^{-1/2}. \qquad (4.13)$$

The function, as prescribed by the techniques mentioned in this report, that measures the similarity of a point x^*, as yet uncategorized, to the category C_1 is the mean-square distance, after transformation, of the unlabeled point to the points of category C_1. This mean-square distance may be written

$$\overline{D^2(x^*,x_k)}^{C_1} = \frac{1}{o_1} \sum_{x_k \in C_1} \sum_{i=1}^{n} \frac{1}{\lambda_i} (x^* a'_i - x_k a'_i)^2, \tag{4.14}$$

where the a_i's are the eigenvectors of A_1 and the λ_i's are the eigenvalues of A_1.

Likewise the mean-square distance of x^* to the points of C_2 is given by

$$\overline{D^2(x^*,x_k)}^{C_2} = \frac{1}{o_2} \sum_{x_k \in C_2} \sum_{i=1}^{n} \frac{1}{\phi_i} (x^* b'_i - x_k b'_i)^2, \tag{4.15}$$

where the b'_is are the eigenvectors and the ϕ_i's are the eigenvalues of A_2.

The decision procedure whereby the point x^* is assigned to C_1 or C_2 consists of observing whether

$$\overline{D^2(x^*,x_k)}^{C_1} - \overline{D^2(x^*,x_k)}^{C_2} \geq K \tag{4.16}$$

or

$$\overline{D^2(x^*,x_k)}^{C_1} - \overline{D^2(x^*,x_k)}^{C_2} < K, \tag{4.17}$$

where K is a number chosen to satisfy some criterion, (e.g., minimization of the false dismissal rate).

It will be shown that the regions defined by (4.16) and (4.17) are the same, except for additive constants, as those defined by (4.9) and (4.10).

We first observe that

$$\overline{D^2(x^*,x_k)}^{C_1} = \frac{1}{c_1} \sum_{x_k \in C_1} \sum_{i=1}^{n} \frac{1}{\lambda_i} \left[(x^* a'_i)^2 - 2\, x^* a'_i x_k a'_i + (x_k a'_i)^2 \right]$$

$$- \sum_{i=1}^{n} \frac{1}{\lambda_i} \left[(x^* a'_i)^2 - 2\, x^* a'_i\, \overline{xa'_i} + \overline{(xa_i)^2} \right], \qquad (4.18)$$

where the averaging takes place over the category C_1.

Next, observe that the variance of the rotated coordinates xa'_i is given by the eigenvalue λ_i. That is,

$$\sigma_i^2 = \overline{(xa'_i)^2} - \overline{xa'_i}^2 = \lambda_i. \qquad (4.19)$$

Adding and subtracting $\overline{xa'_i}^2$ within the brackets of (4.18) and employing (4.19), we obtain

$$\overline{D^2(x^*,x_k)}^{C_1} = \sum_{i=1}^{n} \frac{1}{\lambda_i} \left[\lambda_i + (x^* a'_i - \overline{xa'_i})^2 \right]$$

$$= n + \sum_{i=1}^{n} \frac{1}{\lambda_i} (x^* a'_i - \overline{xa'_i})^2. \qquad (4.20)$$

Similarly, the mean-square distance of x^* to the points of C_2 may be written

$$\overline{D^2(x^*,x_k)}^{C_2} = n + \sum_{i=1}^{n} \frac{1}{\phi_i} (x^* b'_i - \overline{xb'_i})^2, \qquad (4.21)$$

where the averaging takes place over the points of C_2.

If we denote the vector, the components of which are the means of the components of the category C_1 by μ_1, and the corresponding vector for C_2 by μ_2, then the regions (4.16) and (4.17) may be rewritten as

$$\sum_{i=1}^{n} \frac{1}{\lambda_i} (x^* a'_i - \mu_1 a'_i)^2 - \sum_{i=1}^{n} \frac{1}{\phi_i} (x^* b'_i - \mu_2 b'_i)^2 \geq K \qquad (4.16a)$$

$$\sum_{i=1}^{n} \frac{1}{\lambda_i} (x^* a'_i - \mu_1 a'_i)^2 - \sum_{i=1}^{n} \frac{1}{\phi_i} (x^* b'_i - \mu_2 b'_i)^2 < K. \qquad (4.17a)$$

We next observe that

$$\sum_{i=1}^{n} \frac{1}{\lambda_i} (x^* a'_i - \mu_1 a'_i)^2 = (x^* A'_1 \Lambda_1^{-\frac{1}{2}} - \mu_1 A'_1 \Lambda_1^{-\frac{1}{2}})(x^* A'_1 \Lambda_1^{-\frac{1}{2}} - \mu_1 A_1 \Lambda_1^{-\frac{1}{2}})'$$

$$= (x^* - \mu_1) A'_1 \Lambda_1^{-1} A_1 (x^* - \mu_1)' \qquad (4.22)$$

and

$$\sum_{i=1}^{n} \frac{1}{\phi_i} (x^* b'_i - \mu_2 b'_i)^2 = (x^* - \mu_2) A'_2 \Lambda_2^{-1} A_2 (x^* - \mu_2)' \qquad (4.23)$$

It has been observed above that

$$U_1^{-1} = A'_1 \Lambda_1^{-1} A_1$$

and

$$U_2^{-1} = A'_2 \Lambda_2^{-1} A_2,$$

Hence, the regions (4.16a) and (4.17a) may be written

$$(x^* - \mu_1) U_1^{-1} (x^* - \mu_1)' - (x^* - \mu_2) U_2^{-1} (x^* - \mu_2)' \geq K \qquad (4.16b)$$

$$(x^* - \mu_1) U_1^{-1} (x^* - \mu_1)' - (x^* - \mu_2) U_2^{-1} (x^* - \mu_2)' < K. \qquad (4.17b)$$

These regions are the same as those of (4.9) and (4.10) when the constant term K is chosen properly. Various choices of K reflect the criterion employed in the decision procedure. Some of these have been mentioned above (e.g., maximum likelihood, minimax).

The correspondence of the two decision procedures is quite informative. In order to arrive at the decision theory solution, it was necessary to know the density functions of the categories. Knowledge of the a priori probabilities, although not indispensible, was an essential part of the reasoning. The techniques presented in this contract allow a procedure when neither of these factors is known.

5. CATEGORIZATION BY SEPARATION OF CLASSES

5.1 Optimization Criteria

The central concept of the special theory of similarity described in the preceding chapters is that nonidentical events of a common category may be considered close by some method of measuring distance. This measure of distance is placed in evidence by that transformation of the signal space which brings together like events by clustering them most. In this special theory no serious attempt has been made to assure that the metrics which were developed should separate events of different categories.

The purpose of this chapter is to introduce criteria for developing optimum metrics and transformations which not only cluster events of the same class but also separate those which belong to different classes. Consider, for example, the transformation which maximizes the mean-square distance between points which belong to different classes while it minimizes the mean-square distance between points of the same class. The effect of such a transformation is illustrated in Figure 20 where like events have been clustered through minimization of intraset distances and clusters have been separated from each other through the maximization of interset distances. The transformation which accomplishes the stated objectives can be specified by the following problems.

Problem 1

Find the transformation T within a specified class of transformations which maximizes the mean-square interset distance subject to the constraint that the sum of the mean-square interset and intraset distances is held constant.

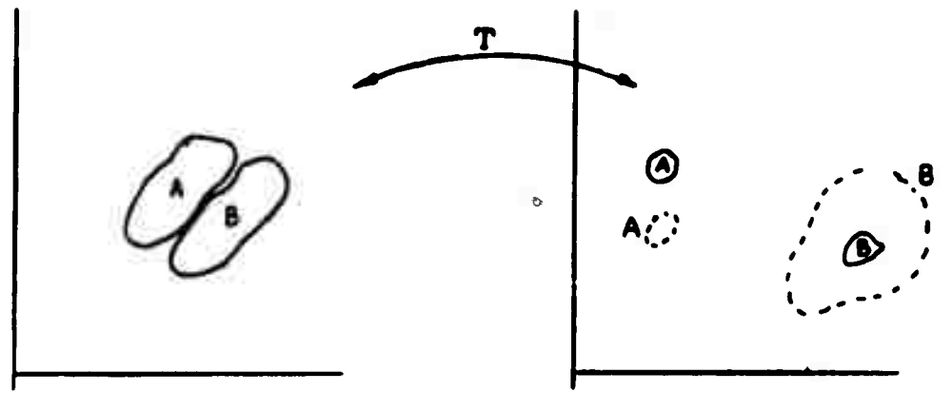


Figure 10. Separation of Classes

Note that for the sake of simplifying the mathematics, the minimization of intraset distances was converted to a constraint on the maximization problem. If interset distances are maximized, and the sum of inter and intraset distances is constant, then it follows that intraset distances are minimized. We may impose the additional constraint that the mean-square intraset distance of each class is equal, thereby avoiding the possible preferential treatment of one class over another. Without the latter constraint the situation indicated with dotted lines in Figure 10 may occur where minimization of the sum of intraset distances may leave one set more clustered than the other.

The above criterion of optimization is given as an illustrative example of how one may convert the desirable objective of separation of classes to a mathematically expressible and solvable problem. Several alternate ways of stating the desired objectives as well as choosing the constraints are possible. For example, the mean-square intraset distance could be

minimized while holding the interset distances constant.

The optimization criterion just discussed suggests a different block diagram for the process of categorization than that shown in Figure 6. Here only a single transformation is developed, resulting in only a single metric with which to measure distance to all of the classes. The classification of an event P is accomplished, as before, by noting to which of the classes the event is most similar. The only difference is that now similarity to each class is measured in the same sense, in the sense exhibited by the transformation which maximally separated events of different categories, on the average.

Problem 2

A second, even more interesting criterion for optimum categorization is the optimization of the classificatory decision on the labeled events. Classificatory decisions are ultimately based on comparing the similarity S (mean-square distance) of the event P with the known events of each class. If P is chosen as any member of Class A, for example, we would like that $S\left|P, \left\{A_m\right\}\right| < S\left|P, \left\{B_m\right\}\right|$, on the average, where $\left\{B_m\right\}$ is the set of known members of any other Class B. Similarly, if P is any member of B, then $S\left|P, \left\{B_m\right\}\right| < S\left|P, \left\{A_m\right\}\right|$. The two desirable requirements are conveniently combined in the statement of the following problem.

Find the metric or transformation of a given class of transformations which maximizes $S\left|P, \left\{B_m\right\}\right| - S\left|P, \left\{A_m\right\}\right|$, on the average, if P belongs to Category A, while requiring that the average of $S\left|P, \left\{A_m\right\}\right| - S\left|P, \left\{B_m\right\}\right|$ for any P contained in Category B is a positive constant. The constraint of this

problem assures that not only points of Category A but also those of B are classified correctly, on the average.*

It is important to note that the above problem is not aimed at maximizing the number of correct decisions. Instead it makes the correct decisions most unequivocal, on the average. It is substantially more difficult to maximize the number of correct classifications. For that purpose a binary function would have to be defined which assumes the more positive of its two values whenever a decision is correct and, conversely, assumes the lower value for incorrect classifications. The sum of this binary function evaluated for each labeled point would have to be maximized. This problem does not lend itself to ready analytical solution; it may be handled, however, by computer methods.

5.2 A Separating Transformation

The particular linear transformation which maximizes the mean-square interset distance while holding the sum of the mean-square inter and intraset distances constant is developed below. Recall that the purpose of this transformation is to separate events of dissimilar categories while clustering those which belong to the same class.

The mean-square distance between the M_1 members of the set $\{F_m\}$ and the M_2 members of the set $\{G_p\}$, after their linear transformation, is given in Equation 5.1, where f_{ms} and g_{ps} are the s^{th} coefficients of the m^{th} and p^{th} members of the sets $\{F_m\}$ and $\{G_p\}$, respectively. For the sake of notational simplicity this mean-square interset distance is denoted by $S|\{F_m\}, \{G_p\}|$ and

* The symmetrical situation where $S|P,\{A_m\}| - S|P,\{B_m\}|$ for PeB is also maximized leads to the same solution.

is the quantity to be maximized by a suitable choice of the linear transformation. The choice of the notation above is intended to signify that the transformation to be found is a function of the two sets.

$$S\left|\{F_m\},\ \{G_p\}\right| = \frac{1}{M_1 M_2}\sum_{m=1}^{M_1}\sum_{p=1}^{M_2}\sum_{n=1}^{N}\left[\sum_{s=1}^{N} w_{ns}\left|f_{ms} - g_{ps}\right|\right]^2 \qquad (5.1)$$

The constraint that the mean-square distance θ between points regardless of the set to which they belong is a constant, is expressed by Equation 5.2, where γ is the coefficient of any point belonging to the union of the sets $\{F\}$ and $\{G\}$, $M_T = \binom{M_1+M_2}{2}$, and $M = M_1+M_2$.

$$\theta = \frac{1}{M_T}\sum_{m=1}^{M}\sum_{p=1}^{M}\sum_{n=1}^{N}\left[\sum_{s=1}^{N} w_{ns}\left|\gamma_{ms} - \gamma_{ps}\right|\right]^2 = \text{constant } K. \qquad (5.2)$$

Both of the above equations may be simplified by expanding the squares as double sums and interchanging the order of summations. Carrying out the indicated operations, we obtain Equations 5.3 and 5.4.

$$S\left|\{F_m\},\ \{G_p\}\right| = \sum_{n=1}^{N}\sum_{s=1}^{N}\sum_{r=1}^{N} w_{ns}\, w_{nr}\, x_{sr}, \qquad (5.3a)$$

where

$$x_{sr} = x_{rs} = \frac{1}{M_1 M_2}\sum_{m=1}^{M_1}\sum_{p=1}^{M_2}\left|f_{ms} - g_{ps}\right|\left|f_{mr} - g_{pr}\right|; \qquad (5.3b)$$

and

$$\Theta = \sum_{n=1}^{N} \sum_{s=1}^{N} \sum_{r=1}^{N} w_{ns}\, w_{nr}\, t_{sr} = K, \tag{5.4a}$$

where

$$t_{sr} = t_{rs} = \frac{1}{M_T} \sum_{m=1}^{M} \sum_{p=1}^{M} \left|\gamma_{ms} - \gamma_{ps}\right| \left|\gamma_{mr} - \gamma_{pr}\right|. \tag{5.4b}$$

The coefficient x_{sr} is the general element of the matrix $[X]$ which is of the form of a covariance matrix and arises from considerations of cross-set distances. The matrix $[T]$ with general coefficient t_{sr}, on the other hand, arises from considerations involving distances between the total number of points of all sets.

We now maximize Equation 5.3, subject to the constraint of Equation 5.4a by the method of Lagrange multipliers. Since dw_{ns} is arbitrary in Equation 5.5, Equation 5.6 must be satisfied.

$$dS - \lambda d\Theta = \sum_{n=1}^{N} \sum_{s=1}^{N} dw_{ns} \left[\sum_{r=1}^{N} w_{nr} \left| x_{sr} - \lambda t_{sr} \right| \right] = 0 \tag{5.5}$$

$$\therefore \sum_{r=1}^{N} w_{nr} \left| x_{sr} - \lambda t_{sr} \right| = 0, \quad \text{for } n=1,2,\ldots,N;\ s=1,2,\ldots,N. \tag{5.6}$$

Equation 5.6 may be written in matrix notation to exhibit the solution in an illuminating way. If we let W_n be a vector with N components $w_{n1} \cdots w_{nN}$, then Equation 5.6 may be written as in Equation 5.7a,

$$\begin{aligned} &W_1\left[X - \lambda T\right] = 0 \\ &\cdots\cdots\cdots \\ &W_n\left[X - \lambda T\right] = 0 \\ &\cdots\cdots\cdots \\ &W_N\left[X - \lambda T\right] = 0 \end{aligned} \qquad (5.7a)$$

By post-multiplying both sides of the equation by T^{-1}, we obtain Equation (5.7b) which is in the form of an eigenvalue problem.

$$\begin{aligned} &W_1\left[XT^{-1} - \lambda I\right] = 0 \\ &W_2\left[XT^{-1} - \lambda I\right] = 0 \\ &\cdots\cdots\cdots\cdots \\ &W_N\left[XT^{-1} - \lambda I\right] = 0 \end{aligned} \qquad (5.7b)$$

Note, that T^{-1} always exists since T is positive definite. Equations 5.7a and b may be satisfied in either ot two ways. Either W_n, the n^{th} row of the linear transformation described by the matrix $\left[W\right]$, is identically zero, or it is an eigenvector of the matrix $\left[XT^{-1}\right]$. We must substitute back into the mean-square interset distance given by Equation 5.3a to find the solution which maximizes S. To facilitate this substitution, we recognize that through matrix notation Equations 5.3a and 5.4a may be written as Equations 5.8 and 5.9, respectively.

$$S\left[\{P_m\}, \{Q_p\}\right] = \sum_{n=1}^{N} W_n [X] W_n^T . \tag{5.8}$$

$$\theta = \sum_{n=1}^{N} W_n [T] W_n^T = K. \tag{5.9}$$

But from Equation 5.7a we see that $W_n X$ may always be replaced by $\lambda W_n T$. Carrying out this substitution in 5.8, we obtain Equation 5.10, where the constraint of 5.9 is also utilized.

$$S\left[\{P_m\}, \{Q_p\}\right] = \lambda \sum_{n=1}^{N} W_n T \; W_n^T = \lambda K \tag{5.10}$$

It is now apparent that the <u>largest</u> eigenvalue of $[X - \lambda T] = 0$ yields the rows of the transformation which maximizes the mean-square interset distance subject to the constraint that the mean-square value of all distances is a constant. The transformation is stated by Equation 5.11, where $W_1 = w_{11}, w_{12}, \cdots, w_{1N}$ = eigenvector corresponding to λ_{max}.

$$[W] = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1N} \\ w_{11} & w_{12} & \cdots & w_{1N} \\ \cdot & \cdot & \cdot & \cdot \\ w_{11} & w_{12} & \cdots & w_{1N} \end{bmatrix}, \tag{5.11}$$

The transformation of the equation above is singular, expressing the fact that the projection of the points along the line of maximum mean-square

interset distance and minimum intraset distance is the only important feature of events that determines their class membership. This is illustrated in Figure 11, where line aa' is in the direction of the first eigenvector of the matrix $[XI^{-1}]$. A point of unknown classification is grouped in Category B because the mean-square difference between its projection on line aa' and the projection of points belonging to set B, $S[P, \{B\}]$, is less than $S[P, \{A\}]$, the corresponding difference with members of set A.

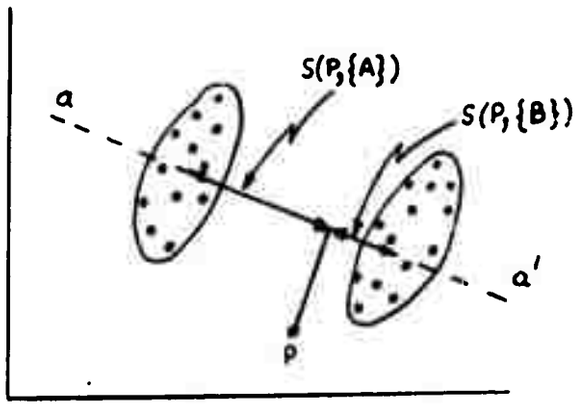


Figure 11 A Singular Class-Separating Transformation

Forcing the separating transformation to be non-singular is possible by the imposition of a different constraint on the maximization. Unfortunately, the mathematical difficulty of imposing non-singularity directly is a formidable task. In general it requires evaluating a determinant, such as the Gramian, and assuring that it does not vanish. In the following discussion, at first a seemingly meaningless constraint will be imposed on the maximization of the mean-square interset distance. After the solution is obtained, it

will be shown that the meaningless constraint can be converted to a constraint which holds the mean-square of all distances constant—the same constraint we used previously.

The mean-square interset distance to be maximized is given by Equation 5.3a which is reproduced here as Equation 5.12.

$$S\left|\{P_m\},\ \{G_p\}\right| = \sum_{n=1}^{N}\sum_{s=1}^{N}\sum_{r=1}^{N} w_{ns}\, w_{nr}\, x_{sr}. \qquad (5.12)$$

The constraint we will impose is that the mean-square length of the projections of all distances between any pair of points onto the directions W_n be fixed, but in general, different constants. This constraint is expressed by Equation 5.13 which differs from the previously used constraint of Equation 5.4 only by fixing coordinate by coordinate the mean-square value of all possible distances between points.

$$\sum_{s=1}^{N}\sum_{r=1}^{N} w_{ns}\, w_{nr}\, t_{sr} = K_n, \text{ for } n=1,2,\ldots,N. \qquad (5.13)$$

Assigning an arbitrary constant λ_n to the differential of each of the above N constraints and using the method of Lagrange multipliers in the maximization of S above, Equation 5.14 is obtained.

$$dS - \sum_{n=1}^{N} \lambda_n\, dK_n = \sum_{n=1}^{N}\sum_{s=1}^{N} dw_{ns}\left[\sum_{r=1}^{N} w_{nr}\left(x_{sr} - \lambda_n t_{sr}\right)\right] = 0. \qquad (5.14)$$

When we make use of the convenient matrix notation employed earlier, we obtain Equation 5.15 which differs significantly from Equation 5.7a, despite the similar appearance of the two equations.

$$\begin{aligned} W_1\left[X - \lambda_1 T\right] &= 0 \\ W_2\left[X - \lambda_2 T\right] &= 0 \\ &\cdots\cdots\cdots \\ W_N\left[X - \lambda_N T\right] &= 0 \end{aligned} \qquad (5.15)$$

The solution of Equation 5.15 states that each row of the linear transformation, W_n, is a <u>different</u> eigenvector of the $\left[X\,T^{-1}\right]$ matrix. The transformation $\left[W\right]$ is therefore orthogonal. Equation 5.16 is a further constraint which converts that of 5.15 to holding the mean-square of all distances constant, and thus accomplishes the aim of this section.

$$K = \sum_{n=1}^{N} K_n \qquad (5.16)$$

Note that before we knew that the rows of the transformation $\left[W\right]$ would be orthogonal, the condition expressed by Equation 5.16 does not fix the total distances. The above procedure resulted in finding the non-singular orthogonal transformation which optimally separates the classes and optimally clusters members of the same class.

We will now compute the mean-square interset distance S of Equation 5.12. To facilitate the computation, S will be written in matrix notation as in Equation 5.17.

When we make use of the convenient matrix notation employed earlier, we obtain Equation 5.15 which differs significantly from Equation 5.7a, despite the similar appearance of the two equations.

$$\begin{aligned} W_1\left[X - \lambda_1 T\right] &= 0 \\ W_2\left[X - \lambda_2 T\right] &= 0 \\ &\cdots\cdots\cdots \\ W_N\left[X - \lambda_N T\right] &= 0 \end{aligned} \qquad (5.15)$$

The solution of Equation 5.15 states that each row of the linear transformation, W_n, is a <u>different</u> eigenvector of the $\left[X\,T^{-1}\right]$ matrix. The transformation $\left[W\right]$ is therefore orthogonal. Equation 5.16 is a further constraint which converts that of 5.15 to holding the mean-square of all distances constant, and thus accomplishes the aim of this section.

$$K = \sum_{n=1}^{N} K_n \qquad (5.16)$$

Note that before we knew that the rows of the transformation $\left[W\right]$ would be orthogonal, the condition expressed by Equation 5.16 does not fix the total distances. The above procedure resulted in finding the non-singular orthogonal transformation which optimally separates the classes and optimally clusters members of the same class.

We will now compute the mean-square interset distance S of Equation 5.12. To facilitate the computation, S will be written in matrix notation as in Equation 5.17.

$$S\left[\{F_m\}, \{G_p\}\right] = \sum_{n=1}^{N} W_n X W_n^T \tag{5.17}$$

From Equation 5.15 it is seen, however, that if S is maximum, $W_n X$ may be replaced with $\lambda_n W_n T$ to obtain Equation 5.18, where $W_n T W_n^T = K_n$

$$S_{max}\left[\{F_m\}, \{G_p\}\right] = \sum_{n=1}^{N} \lambda_n W_n T W_n^T, \tag{5.18}$$

from Equation 5.13 (in matrix notation). Equation 5.19 is thus obtained. It is now readily seen, with reference to Equation 5.10, that the upper bound on the mean-square interset distance is achieved by the singular transformation discussed earlier, and we pay for forcing the transformation to be non-singular by achieving only a reduced separability of classes.

$$S_{max}\left[\{F_m\}, \{G_p\}\right] = \sum_{n=1}^{N} \lambda_n K_n. \tag{5.19}$$

Before leaving the discussion of class-separating transformations, a few important facts must be pointed out. A simple formal replacement of the matrices X and T by other suitably chosen matrices yields the solution of many interesting and useful problems. It is not the purpose of the following remarks to catalog the problems solved by the formal solution previously obtained; yet some deserve mention because of their importance. It may be readily verified, for instance, that replacing T by I is equivalent to maximizing the between-set distances, subject to the condition that the volume

of the space is a constant. The tranformation which accomplishes this is orthogonal with rows equal to different eigenvectors of the matrix X. This is a physically obvious result, of course, since the eigenvectors of X are the set of orthogonal directions along which interset distances are maximized, on the average. A figure which would illustrate the result is very similar to Figure 11.

Another replacement which must be singled out is the substitution of the matrix L for T, where L is the covariance matrix associated with all intraset distances (distances among like events). Eigenvectors of $[X - \lambda L]$ form rows of the transformation which maximizes interset distances while holding intraset distances constant. This problem is essentially the same as the maximization of interset distances while holding the sum of inter and intraset distances constant, yet the relative separation of sets achieved by the two transformations is different. The difference may be exhibited by computing the ratio of the mean-square separation of sets to the mean clustering of elements within the same set, as measured by the mean-square intraset distance. It may be concluded, therefore, that the constraint employed in the maximization of interset distances does have an influence on the degree of separation achieved between sets.

Throughout this chapter the class-separating transformations were developed by reference to the existence of only two sets, $\{F_m\}$ and $\{G_p\}$. The results obtained by these methods are more general, however, because they apply directly to the separation of an arbitrary number of sets. For instance, in the maximization of the mean-square interset distance, there is no reason why the matrix X should involve interset distances between only two sets. An

arbitrary number of sets may be involved, and the interset distances are simply all those distances measured between two points not in the same set. Similar arguments are valid for all the other matrices involved. The only precaution that must be taken concerns the possible use of additional constraints specifying preferential or nonpreferential treatment of classes. These additional constraints may take the form of requiring, for instance, that the mean-square intraset distance of all sets be equal or be related to each other by some constants. Aside from these minor matters, the results apply to the separation of any number of classes.

5.3 Maximization of Correct Classifications

The correct classification of points of the set F are made more unequivocal by the linear transformation which makes any event F_n of set F more similar to members of F, on the average, than to those of another set G. One of the ways in which the average unequivocalness of correct classificatory decisions may be stated mathematically is to require that a numerical value associated with the quality of a decision be maximized, on the average. Of the several quantitative measures of the quality of a decision which may be defined, one that readily lends itself to mathematical treatment is given in Equation 5.20. The difference in the similarity between a point P and each of the two sets, F and G, is a quantity Q which is larger if the decision regarding the classification of P is more unequivocal.

$$S\left|P, \left\{G_m\right\}\right| - S\left|P, \left\{F_m\right\}\right| = Q \qquad (5.20)$$

Since decisions in previous chapters were based on the comparison of Q with a suitable threshold value (such as zero), we now wish to find that linear transformation which maximizes Q, on the average, whenever Q is to be positive. If P is a member of the set F, then P is closer to F than to G and thus Q is to be positive. The maximization of Q for P∈F results in maximizing the margin with which correct decisions are made, on the average. The foregoing maximization is stated in Equation 5.21 subject to the constraint expressed by Equation 5.22. The latter simply states that if P∈G, the average decision is still correct, as measured by the margin K*

$$\overline{S\left(P_n, \{G_p\}\right) - S\left(P_n, \{F_m\}\right)}^{n} = \bar{Q} = \text{maximum, subject to} \tag{5.21}$$

$$\overline{S\left(G_n, \{F_m\}\right) - S\left(G_n, \{G_p\}\right)}^{n} = \bar{K} = \text{constant} > 0. \tag{5.22}$$

Utilizing previously obtained results, the above equations are readily solved for the optimum linear transformation. Rewriting the first term of Equation 5.21, we note that it expresses the mean-square interset distance between sets F and G and may be written as in Equation 5.23, where Equation 5.1 and the simplifying notation of 5.3 are employed.

$$\overline{S\left(P_n, \{G_p\}\right)}^{n} = S\left(\{F_n\}, \{G_p\}\right) = \frac{1}{M_1 M_2} \sum_{m=1}^{M_1} \sum_{p=1}^{M_2} \sum_{n=1}^{N} \left[\sum_{s=1}^{N} w_{ns} \left(f_{ms} - g_{ps} \right) \right]^2 \tag{5.23a}$$

*Maximization of $\bar{Q} \div \bar{K}$ has the same solution.

$$\overline{S\left(P_n, \{G_D\}\right)}^{n} = \sum_{n=1}^{N}\sum_{s=1}^{N}\sum_{r=1}^{N} w_{ns}\, w_{nr}\, x_{sr} \,. \tag{5.23b}$$

The second term of Equation 5.21 is the mean-square intraset distance of set P and may be expressed as in Equation 5.24. The argument of the covariance coefficient $u_{sr}(P)$ signifies that it is a covariance of elements of the set P.

$$\overline{S\left(P_n, \{P_m\}\right)}^{n} = S\left(\{P_n\}, \{P_m\}\right) = \frac{1}{(M_1-1)M_1} \sum_{p=1}^{M_1}\sum_{m=1}^{M_1}\sum_{n=1}^{N}\left[\sum_{s=1}^{N} w_{ns}\left(f_{ps}-f_{ms}\right)\right]^2 \tag{5.24a}$$

$$\overline{S\left(P_n, \{P_m\}\right)}^{n} = \frac{2M_1}{M_1-1} \sum_{n=1}^{N}\sum_{s=1}^{N}\sum_{r=1}^{N} w_{ns}\, w_{nr}\, u_{sr}(P). \tag{5.24b}$$

Similarly, the first term of Equation 5.22 is the mean-square interset distance, and the second term is the intraset distance of set G. The maximization problem can thus be restated by Equation 5.25a and b.

$$\text{Maximize } \bar{Q} = \sum_{n=1}^{N}\sum_{s=1}^{N}\sum_{r=1}^{N} w_{ns}\, w_{nr}\left[x_{sr} - \frac{2M_1}{M_1-1}\, u_{sr}(P)\right], \tag{5.25a}$$

$$\text{subject to } \bar{K} = \sum_{n=1}^{N}\sum_{s=1}^{N}\sum_{r=1}^{N} w_{ns}\, w_{nr}\left[x_{sr} - \frac{2M_2}{M_2-1}\, u_{sr}(G)\right]. \tag{5.25b}$$

Following the methods used earlier, the solution of the above problem may be written down by inspection.

$$d\bar{Q} - \lambda d\bar{R} = \sum_{n=1}^{N} \sum_{s=1}^{N} dw_{ns} \left[\sum_{r=1}^{N} w_{nr} \left\{ x_{sr} - \frac{2M_1}{M_1 - 1} u_{sr}(F) - \lambda \left[x_{sr} - \frac{2M_2}{M_2 - 1} u_{sr}(G) \right] \right\} \right] \quad (5.26)$$

From 5.26 it follows that Equation 5.27a must hold, where α_{sr} and β_{sr} are given by Equations 5.27b and c.

$$\sum_{r=1}^{N} w_{nr} \left[\alpha_{sr} - \lambda \beta_{sr} \right] = 0, \text{ for } n=1,2,\ldots,N \text{ and } s=1,2,\ldots,N \quad (5.27a)$$

$$\alpha_{sr} = x_{sr} - \frac{2M_1}{M_1 - 1} u_{sr}(F) \quad (5.27b)$$

$$\beta_{sr} = x_{sr} - \frac{2M_2}{M_2 - 1} u_{sr}(G) \quad (5.27c)$$

By reference to earlier results, such as those expressed by Equation 5.6, the transformation whose coefficients w_{ns} satisfy an equation of the form above, is the solution of the eigenvalue problem of Equation 5.28, where W_n is a row of the matrix expressing the linear transformation.

$$\begin{aligned} W_1 \left[\alpha - \lambda \beta \right] &= 0 \\ W_2 \left[\alpha - \lambda \beta \right] &= 0 \\ \cdots\cdots\cdots & \\ W_N \left[\alpha - \lambda \beta \right] &= 0 \end{aligned} \quad (5.28)$$

Analogous to the arguments used in the previous section, the above solution yields a singular transformation. Forcing the transformation to be non-

singular, in the manner already outlined, results in the optimum transformation as an orthogonal transformation, where each row of the matrix $[W]$ is an eigenvector of $[\alpha - \lambda\beta] = 0$. Furthermore, it is readily shown that the solution so obtained indeed maximizes $\bar{Q}$.

It is interesting to note that the maximization of the average correct classifications can be considered as the maximization of the difference between inter and intraset distances. This alternate statement of the problem may be exhibited by the addition of Equation 5.25b to 5.25a

$$\bar{Q} + K = \sum_{n=1}^{N} \sum_{s=1}^{N} \sum_{r=1}^{N} w_{ns}\, w_{nr} \left[2x_{sr} - \left\{ \frac{2M_1}{M_1 - 1}\, u_{sr}(F) + \frac{2M_2}{M_2 - 1}\, u_{sr}(G) \right\} \right] \quad . \quad (5.29)$$

But the expression within the braces is simply the covariance ℓ_{sr} associated with all intraset distances. Since K is a constant, the maximization of Equation 5.29 is equivalent to the maximization of $\bar{Q}$.

In summing up the results of this chapter, it is seen that the problem of learning to measure similarity to events of a common category, while profiting from knowledge of nonmembers of the same category, may be treated as a maximization or minimization problem. A metric or a linear transformation is found from a class of metrics or transformations which solves mathematical problems which express the desire not only to cluster events known to belong to the same category, but also to separate those which belong to different categories. Within the restricted class of metrics or transformations considered in this chapter, the solutions are in the form of eigenvalue problems which emphasize features that examples of a category have in common, and which at the same time differ from features of other categories.

7. CONCLUSIONS AND RECOMMENDATIONS

The application of pattern recognition to missile detection and decoy diacrimination was investigated. As a result of these investigations, a mathematical theory was developed to achieve the optimum combination of measurable or computable target properties to achieve the separation of target types, based on their observable properties. The input to the theory is the set of discrimination techniques and useful target-signature parameters. The theory operates on these inputs to determine their optimum method of combination and thus determine what the essential properties of targets are that allow their recognition. An important attribute of the theory is that it does not become obsolete as new techniques are developed and break-throughs are achieved in methods of distinguishing one type of target or threat from snother. The achievement of the present theory is that it states how techniques developed by others are to be combined.

The recommendation for further work, in part, is to develop the theory — which has thus far shown promising results — and to explore the already known avenues of research opened up by current investigations under the present program. Simultaneously with further theoretical developments, the already successfully tested methods should be applied to actual missile data. Detailed recommendations have been outlined in a proposal.

In the course of work performed during the present contractual effort, and as a result of the investigations carried out during this effort, certain facts relating to Ballistic Missile Defense, and in particular, to the problems of Missile Detection and Decoy Discrimination, became apparent. These facts are listed below and are substantiated in the body of this report.

a) It may be inferred from the results of the present study that present computers are entirely adequate in size and in speed of operation to handle the job of Missile Detection and Decoy Discrimination in Ballistic Missile Defense. Neither the complexity nor the number of calculations are outside the capabilities of presently available computers.

b) The instrumentation of the theoretical techniques developed during the present effort would serve a two-fold purpose. On the one hand, it is the working system with which missile detection and decoy discrimination could be carried out in the field: on the other hand, it would serve as a research tool with which both the usefulness of new sensors and recognition techniques can be evaluated, and the direction of desirable further data collection efforts pinpointed.

c) It is felt that the system suggested in this report is both the final system which performs missile detection and decoy discrimination in the field, as well as the tool with which to discover how the job is to be done and what the missile signatures are which permit discrimination between missiles and other targets. For this reason, it appears that, stemming from this dual use, the proposed system will accomplish a significant saving in the developmental time of a ballistic-missile defense system.

d) It is generally realized that no single measurable or computable parameter will be adequate to perform missile detection and decoy discrimination with sufficient reliability. It is believed that a combination of techniques will be necessary to perform these tasks. While such notions are generally entertained, the precise method of combination of the various and basically different techniques remains unsolved. In this

report we can specify exactly how the different present and future techniques which are useful in one situation or another can be combined to yield optimal discriminatory decisions. Furthermore, the manner of combination of techniques will not require alteration either conceptually or in the hardware instrumentation, if new and improved techniques become available.

e) An important conclusion of the present study effort is that the best utilization of a combination of techniques in a single decision process is achieved only through the simultaneous collection of data. It is absolutely necessary that simultaneously collected measurements by all sensors and data processing systems be available if the optimum combination of techniques is to be found.

f) Since a technique of data processing which utilizes the combination of detection and discrimination techniques transcending several different kinds of primary sensors will be used, such a system must face the problem of lack of uniformity of data storage. It is one of the conclusions of the present study that uniformity of data storage should be a desirable objective from the very beginning, lest the problem take on insurmountable proportions later on. The requirement should be established to maintain adherence to a mutually satisfactory data storage format.

BIBLIOGRAPHY

Ashby, W. Ross. Design for a Brain, John Wiley and Sons, Inc., New York (Chapman and Hall, Ltd., London, England) (1952)

Bar-Hillel, Y. "Can Translation be Mechanized?" American Scientist, Vol. 42, pp. 248-260 (April 1954)

Bar-Hillel, Y. "Linguistic Problems Connected with Machine Translation," British Journal for the Philosophy of Science, Vol. 20, No. 3, pp. 217-225 (July 1953)

Bar-Hillel, Y. "The Present State of Research on Mechanical Translation," American Documentation, Vol. II, No. 4 (1952)

Bernstein, A., et al. "A Chess Playing Program for the IBM 704," Proceedings of the Western Joint Computer Conference, pp. 157-159 (May 6-8, 1958)

Bomba, J. S. "Alpha-Numeric Character Recognition Using Local Operations," Paper Presented at Eastern Joint Computer Conference (3 December 1959)

Bremer, R. W. "A Checklist of Intelligence for Programming Systems," Communications of the Association for Computing Machinery, Vol. 2, pp. 8-13 (March 1959)

Carr, J. W. "Recursive Subscripting Compilers and List-Type Memories," Communications of the Association for Computing Machinery, Vol. 2, pp. 4-6 (February 1959)

Chow, C. K. "An Optimum Character Recognition System Using Decision Functions," IRE Transactions on Electronic Computers, Vol. EC-6, pp. 247-254 (December 1957)

Clark, W. A., and Farley, B. G. "Generalizations of Pattern Recognition in a Self-Organizing System," Proceedings of the Western Joint Computer Conference (1955)

David, Jr. E. E. "Artificial Auditory Recognition in Telephony," IBM Journal of Research and Development, Vol. 2, No. 4 (1958)

David, Jr., E. E., and McDonald, H.S. "A Bit-Squeezing Technique Applied to Speech Signals," IRE Convention Record, pp. 148-152 (1956)

Denes, P. "The Design and Operation of the Mechanical Speech Recognizer," Journal of British IRE, Vol. 19, pp. 219-229 and Discussion pp. 230-234 (April 1959)

Dimond, T. L. "Devices for Reading Handwritten Characters," Proceedings of Eastern Joint Computer Conference, pp. 232-237 (1957)

Dinneen, G. P. "Programming Pattern Recognition," Proceedings of the Western Joint Computer Conference (March 1955)

Dunker, K. "On Problem Solving," Psychological Monographs, Vol. 58, No. 270 (1945)

Evey, R. J. "Use of a Computer to Design Character Recognition Logic," Paper presented at Eastern Joint Computer Conference (3 December 1959)

Feldman, J. "A Theory of Binary Choice Behavior," CIP Working Paper No. 12, Carnegie Institute of Technology (May 1958)

Flores, I. "An Optimum Character Recognition System Using Decision Functions," IRE Transactions on Electronic Computers, Vol. EC-7, p. 180 (June 1958)

Friedberg, R. M. "A Learning Machine," Part I, IBM Journal of Research and Development, Vol. 2 (January 1958)

Fucks, W. "On Mathematical Analysis of Style," Biometrika, Vol. 39, p. 122 (1952)

Galanter, Eugene H. "The Behavior of Thought," Paper presented at the American Psychological Association Meeting in Chicago (1956)

Gardner, M. Logic Machines and Diagrams, McGraw-Hill Book Company, Inc., New York (1958)

Gelernter, H. L , and Rochester, N. "Intelligent Behavior in Problem-Solving Machines," IBM Journal of Research and Development, Vol. 2, No. 4 (October 1958)

Gentzen, Gerhard. "Untersuchungen uber das logische Schliessem," Mathematische Zeitschrift, Vol. 39, pp. 176-210 and 405-431 (1934)

Glantz, H. T. "On the Recognition of Information with a Digital Computer," Journal of the Association for Computing Machinery (April 1957)

Gold, B. "Machine Recognition of Hand-Sent Morse Code," IRE Transactions on Information Theory, Vol. IT-5, pp. 17-24 (March 1950)

Greanias, E. C., et al. "Design of Logics for Recognition of Printed Characters by Simulation," IBM Journal of Research and Development, Vol. 1, pp. 8-18 (January 1957)

Grimsdale, R. L., et al. "A System for the Automatic Recognition of Patterns," Journal of the Institution of Electrical Engineers, London, Vol. 106, Part B, pp. 210-221 (March 1959)

Harris, Robert T., and Jarrett, J. L. Language and Informal Logic, Longmans Green & Co., Inc., New York (1956)

Hebb. D. O. The Organization of Behavior, John Wiley and Sons, Inc., New York (Chapman and Hall, Ltd. London) (1949)

Hilgard, E. Theories of Learning, Second Edition, Appleton-Century-Crofts, Inc., New York (1956)

Hovland, C. I. "A Communication Analysis of Concept Learning," Psychological Review, Vol. 59 (1952)

Humphrey, G. Thinking, John Wiley and Sons, Inc. New York (1951)

Ianov, IU.I. "On the Equivalence and Transformation of Program Schemes," Doklady Akademii Nauk S.S.S.R., Vol. 113, No. 1 (1957)

Ianov, IU.I. "On Matrix Program Schemes," Doklady Akademii Nauk S.S.S.R., Vol. 113, No. 2 (1957). Also published in Communications of the Association of Computer Machinery, Vol. 1, No. 12 (December 1958)

Kirsch, L. C., et al. "Experiments in Processing Pictorial Information with a Digital Computer," Proceedings of the Eastern Joint Computer Conference, pp. 221-230 (December 1957)

Kister, J., et al. "Experiments in Chess," Journal of the Association for Computing Machinery, Vol. 4, p. 2 (April 1957)

Kramer, H. P., and Mathews, M. V. "A Linear Coding for Transmitting a Set of Correlated Signals," IRE Transactions on Information Theory, Vol. IT-2 (September 1956). Paper presented at Symposium on Information Theory held at MIT, September 10-12 1956.

Kretzmer, E. R. "Reduced Alphabet Representation of Television Signals," IRE Convention Record, Information Theory section, Part 4, p. 140 (1956)

Lambek, J. "The Mathematics of Sentence Structure" American Mathematical Monthly, Vol. 65 (3 March 1958)

Lashley, K. S. Cerebral Mechanism in Behavior, John Wiley and Sons, Inc., New York (1951)

Latil, de P. Thinking by Machine, Houghton-Mifflin Co., Boston (1956)

Levin, K. Principles of Topological Psychology, McGraw-Hill Book Company, Inc., New York (1936)

Locke, W.N., and Booth, A.D. Machine Translation of Languages, John Wiley and Sons, Inc., New York (1955)

Luchins, A.S. "Mechanization in Problem Solving," Psychological Monographs, Vol. 54, No. 6 (1942)

Luhn, H. "The Automatic Creation of Literature Abstracts," IBM Journal of Research and Development, Vol. 2, No. 2 (April 1958)

Mattson, Richard L. "A Self Organizing Logical System," paper presented at the Eastern Joint Computer Conference (3 December 1959)

McCarthy, J. in C. E. Shannon and J. McCarthy (editors). Automata Studies, Princeton University Press, p. 175 (1956)

McCulloch, W.S., and Pitts, W. "A Logical Calculus of the Ideas Imminent in Nervous Activity," Bulletin of Mathematical Biophysics, Vol. 9, p. 127 (1947)

McCulloch, W.S., et.al. "Symposium on the Design of Machines to Simulate the Behavior of the Human Brain," IRE Transactions on Electronic Computers, Vol. EC-5, No. 4 (December 1956)

"The Mechanization of Thought Processes," Computer Bulletin, Vol. 2, pp. 92,93 (April-May 1959)

Miller, G.A. Language and Communication, McGraw-Hill Book Company, Inc., New York (1951)

Miller, G.A. "The Magical Number Seven," Psychological Review, Vol. 63 (1956)

Miller, G. A., and Selfridge, J.A. "Verbal Context and the Recall of Meaningful Material," American Journal of Psychology, Vol. 63 (1956)

Minsky, Marvin L. "Exploration Systems and Syntactic Processes," (Unpublished Report) Summer Research Project on Artificial Intelligence, Dartmouth College, New Hampshire(1956)

Minsky, Marvin L. "Heuristic Aspects of the Artificial Intelligence Problem," Group Report 34-35, Lincoln Lab., MIT, pp. I-1-I-24 (17 December 1956)

Moore, O. K., and Anderson, S. B. "Modern Logic and Tasks for Experiments on Problem Solving Behavior," Journal of Psychology, Vol. 38, 151-160 (1954)

More, Jr., Trenchard. "Deductive Logic for Automata," (Unpublished) Master's Thesis, Massachusetts Institute of Technology (Elec. Engrg. Dept) (1957)

Morris, Charles. Signs, Language and Behavior, Prentice-Hall, Inc., New York (1946)

Neumann, J. von. "The General and Logical Theory of Automata," in Jeffress (editor), Cerebral Mechanism in Behavior, John Wiley and Sons, Inc., New York (1951)

Neumann, J. von. Theory of Games and Economic Behavior, Princeton University Press (1947)

Newell, A. "The Chess Machine," Proceedings of the Western Joint Computer Conference (March 1955)

Newell, A; Shaw, J. C., and Simon, H. A. "Chess-Playing Programs and the Problem of Complexity," IBM Journal of Research and Development, Vol. 2, No. 4, pp. 320-335 (October 1958)

Newell, A., Shaw, J. C., and Simon, H. A. "Elements of a Theory of Human Problem Solving," the Rand Corporation Report No. P-971, Santa Monica, Calif. (4 March 1957)

Newell, A., Shaw, J. C., and Simon, H. A. "The Elements of a Theory of Human Problem Solving," Psychological Review, Vol. 65 (March 1958)

Newell, A., Shaw. J. C., and Simon, H. A. "Empirical Exploration of the Logic Theory Machine." Proceedings of the Western Joint Computer Conference (February 1957)

Newell, A., Shaw, J. C., and Simon, H. A. "Empirical Exploration of the Logic Theory Machine," (revised), the Rand Corporation Report No. P-951 Santa Monica, Calif. (March 14, 1957)

Newell, A., Shaw, J. C., and Simon, H. A. "General Problem Solving Program," CIP Working Paper No. 7, Carnegie Institute of Technology (December 1957)

Newell, A., Shaw, J. C., and Simon, H. A. "The Processes of Creative Thinking," The Rand Corporation Report No. P-1320 (August 1958)

Newell, A., Shaw, J. C., and Simon, H. A. "Report on a General Problem Solving Program," The Rand Corporation Report No. P-1584 (January 1959)

Newell, A., and Shaw, J. C. "Programming the Logic Theory Machine," Proceedings of the Western Joint Computer Conference (February 1957)

Newell, A., and Shaw, J. C. "Programming the Logic Theory Machine" (revised), The Rand Corporation Report No. P-954, Santa Monica, Calif. (28 February 1957)

Newell, A., and Simon, H. A. "Current Developments in Complex Information Processing," The Rand Corporation Report No. P-850, Santa Monica, Calif. (May 1, 1956)

Newell, A., and Simon, H. A. "The Logic Theory Machine," IRE Transactions on Information Theory, Vol. IT-2, No. 2, pp. 61-79 (September 1956)

Newell, A., and Simon, H. A. "The Logic Theory Machine. A Complex Information Processing System" (revised", The Rand Corporation Report No. P-868, Santa Monica, Calif. (12 July 1956)

Oettinger, A. G. "Simple Learning by a Digital Computer," Proceedings of the Association for Computing Machinery (September 1952)

Perry, J. W., Kent, A. and Berry, N. M. Machine Literature Searching, Interscience Publishers, Inc., New York (1956)

Pitts, W., and McCulloch, W. S. "How We Know Universals, the Perception of Auditory and Visual Form," Bulletin Mathematical Biophysics, Vol. 9 (1947)

Polya, G. How to Solve It, Princeton University Press (1945)

Polya, G. Mathematics and Plausible Reasoning, Vols. I and II, Princeton University Press (1954)

Rapaport, D. The Organization and Pathology of Thought, Columbia University Press, New York (1951)

Rochester, N., et al. "Tests on a Cell Assembly Theory of the Action of the Brain Using a Large Digital Computer," IRE Transactions on Information Theory, IT-2, No. 3 (September 1956)

Rosenblatt, F. "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain," Psychological Review, Vol. 65, No. 6 (November 1958)

Rosenblatt, F. "The Perceptron, A Theory of Statistical Separability in Cognitive Systems," Cornell Aeronautical Laboratory, Project PARA, Report No. VG-1196-G-1 (January 1958)

Selfridge, Oliver G. "Pandemonium: A Paradigm for Learning," Proceedings of the Symposium on Mechanization of Thought Processes, held at the National Physical Laboratory, Teddington, Middlesex, England (24-27 November, 1958)

Selfridge, Oliver G. "Pattern Recognition and Learning," Symposium on Information Theory, London, England (1955). Preprinted Group Report 34-43, Lincoln Laboratory of Massachusetts Institute of Technology (July 20, 1955)

Selfridge, Oliver G. "Pattern Recognition and Modern Computers," Proceedings of the Western Joint Computer Conference, pp. 91-93, (March 1955)

Selfridge, Oliver G., et al. "Pattern Recognition and Reading by Machine," Proceedings of the Eastern Joint Computer Conference (December 3, 1959)

Shannon, Claude E. "Communication Theory of Secrecy Systems," Bell System Technical Journal, Vol. 28, pp. 656-715 (1949)

Shannon, Claude E. "Computers and Automata," Proceedings of the IRE, Vol. 41 (March 1950)

Shannon, Claude E. "A Mathematical Theory of Communication," Bell System Technical Journal, Vo.. 27, pp. 379-423 (October 1948)

Shannon, Claude E. "Prediction and Entropy of Printed English" Bell System Technical Journal, Vol. 30. pp. 50-64 (1951)

Shannon, Claude E. "Programming a Computer for Playing Chess," The Philosophical Magazine, Vol. 41 (March 1950)

Shannon, Claude E. "The Rate of Approach to Ideal Coding," IRE National Convention Record, Part 4, (Abstract pages only) (1955)

Shannon, C. E., in J. McCarthy (editor). "Automata Studies," Annals of Mathematics Studies, No. 4, Princeton (1956)

Shaw, J. C., et al. "A Command Structure for Complex Information Processing," Proceedings of the Western Joint Computer Conference (May 1958)

Simon, H. A. "A Behavioral Model of Rational Choice," The Quarterly Journal of Economics, Vol. 69 (February 1955)

Simon, H. A. "Rational Choice and the Difficulty of the Environment," Psychological Review, Vol. 63 (March 1956)

Simon, H. A., and Newell, A. "Models: Thier Uses and Limitations," in White (editor), The State of Social Sciences, University of Chicago (1956)

Simons, Leo. "New Axiomatizations of S3 and S4," Journal of Symbolic Logic, Vol. 18, No. 4, pp. 309-316 (1953)

Solomonoff, R. J. "An Inductive Inference Machine (privately circulated report) (August 14, 1956); IRE National Convention Record, Vol. 5, Part 2, pp. 56-62 (1957); Annals of Mathematical Studies, No. 34, Princeton (1956)

Solomonoff, R. J. "A New Method for Discovering the Grammars of Phase Structure Languages," AFOSR TN-59-110, under Contract No. AF49(638)-376 (April 1959) (ASTIA AD No. 210 390)

Solomonoff, R. J. "The Mechanization of Linguistic Learning," AFOSR-TN-246 under Contract No. AF49(638)-376 (April 1959) (ASTIA AD No. 212 226)

Steinbuch, K. "Automatic Speech Recognition," Nachrichtentechnische Zeitschrift, Vol. 11, pp. 446, 454 (September 1958)

Strachey, C. S. "Logical or Non-Mathematical Programs," Proceedings of the Association for Computing Machinery (September 1952)

Taylor, W. K. "Pattern Recognition by Means of Automatic Analogue Apparatus," Proceedings of the IRE, (London), Vol. 106, Part B, pp. 198-209 (March 1959)

Tersoff, A. I. "Electronic Reader Sorts Mail," Electronic Industries, pp. 56-60 (July 1958)

Turing, A. M. "Can a Machine Think?" in J. R. Newman, The World of Mathematics, Vol. 4, Simon and Shuster, Inc., New York (1956)

Turing, A. M. "On Computable Numbers," Proceedings of the London Mathematical Society, Series 2, Vol. 42 (1936-37). See also a correction, Ibid, Vol. 43 (1937)

Unger, S. H. "A Computer Oriented Toward Spatial Problems," Proceedings of the IRE, Vol. 46, pp. 1744-1750 (October 1958)

Unger, S. H. "Pattern Detection and Recognition," Proceedings of the IRE, Vol. 47, No. 10, p. 1737 (October 1959)

Uttley, A. M. "The Classification of Signals in the Nervous System," Radar Research Establishment Memorandum 1047, Great Malvern, England (1954). Also published in the Journal of Electroencephalography and Clinical Neurophysiology, Vol. 6, p. 479 (1954)

Uttley, A. M. "The Probability of Neural Connections," Radar Research Establishment Memorandum 1048, Great Malvern, England (1954)

Yngve, V. "Programming Language for Mechanical Translation," Mechanical Translation, Vol. 5, No. 1 (July 1958)

APPENDIX A

The Solution of Eigenvalue Problems

The frequency with which eigenvalue problems occur in classificatory analysis and the difficulty with which they are solved warrants the careful examination of the available methods of solution. The slowest link in the computations involved in the process of forming category membership measuring functions is the solution of eigenvalue problems. It takes a considerable length of time for even a fast computer to solve for the eigenvalues and vectors of a large matrix. This limits the speed with which the influence of a new event is felt on the recognition process. In order that machine learning be carried out in essentially "real time", it is necessary to search for a physical phenomenon or a natural process which is the solution of an eigenvalue problem. The natural phenomenon must have enough controllable parameters to allow the setting up of an arbitrary positive definite symmetric matrix. The objective of this Appendix is to focus attention on the importance of finding such a natural phenomenon and to give an example which—although not completely general, as we shall see, nor a practical as some would like—does demonstrate the feasibility of solving eigenvalue problems very rapidly.

Consider the two-loop lossless network of Figure A-1 which is excited with a voltage source $\underline{e}$ at its input. Letting the complex frequency be λ and the reciprocal capacitance (susceptance) values be called $\underline{S}$, the loop equations of the network may be written as in Equation A.1.

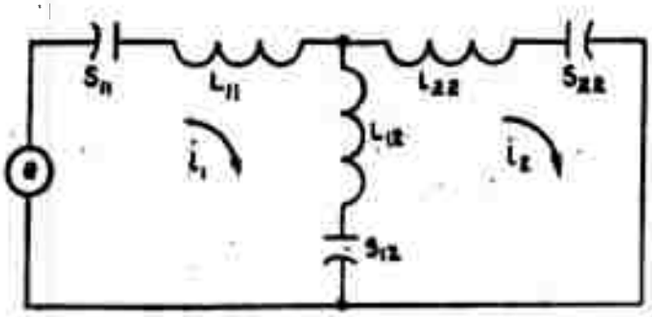


Figure A-1. Two-Loop Lossless Network

$$e_1 = \left[\lambda\left(L_{11}+L_{12}\right) + \frac{S_{11}+S_{12}}{\lambda}\right] i_1 - \left[\lambda L_{12} + \frac{S_{12}}{\lambda}\right] i_2 \tag{A.1}$$

$$0 = -\left[\lambda L_{12} + \frac{S_{12}}{\lambda}\right] i_1 + \left[\lambda\left(L_{22}+L_{12}\right) + \frac{\left(S_{22}+S_{12}\right)}{\lambda}\right] i_2$$

Multiplying both sides of the equation by λ and writing it in matrix notation, we obtain Equation A.2, where $\underline{e}$ and $\underline{i}$ are vectors of the voltage excitations in the loops and loop currents, respectively.

$$\lambda \underline{e} = \underline{i}\left(\lambda^2[L] + [S]\right) \tag{A.2}$$

The matrices $[L]$ and $[S]$ are given in Equation A.3.

$$[L] = \begin{bmatrix} L_{11}+L_{12} & -L_{12} \\ -L_{12} & L_{22}+L_{12} \end{bmatrix}; \quad [S] = \begin{bmatrix} S_{11}+S_{12} & -S_{12} \\ -S_{12} & S_{22}+S_{12} \end{bmatrix} \tag{A.3}$$

If the input is short-circuited and the vector $\underline{e}$ is zero, any non-zero current that flows in the network must do so at frequencies that satisfy Equation A.4, where use is made of the knowledge that a lossless network must oscillate at pure imaginary frequencies $\lambda = j\omega$. The resulting equation is an eigenvalue problem of the same type encountered throughout in this volume.

$$\left[\lambda^2 L + C\right] = 0 = \left[C - \omega^2 L\right] \tag{A.4}$$

The matrix $\left[L\right]$ is a completely arbitrary, symmetric, positive definite matrix whose coefficients each are controlled by (at most) two circuit elements. The matrix $\left[C\right]$ is also symmetric and positive definite, but its elements which are off the principal diagonal must be negative or zero. This does not have to be the case in the $\left[L\right]$ matrix for a negative mutual inductance is quite realizable. Note, however, that if the mutual capacitance is short-circuited, and the other capacitors are made equal (for convenience let them be unity), then the natural frequencies of oscillation of the short-circuited network satisfy the eigenvalue problem of Equation A.5.

$$0 = \left[L - \frac{1}{\omega^2} I\right] \tag{A.5}$$

The most general two-loop network corresponding to this equation is shown in Figure A-2, where a transformer replaces the mutual inductances.

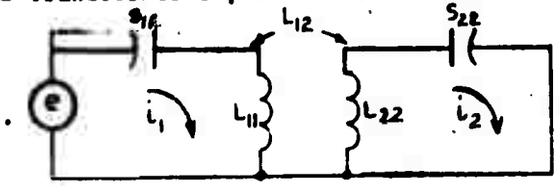


Figure A-2. Network Solution of an Eigenvalue Problem

The eivenvalues are the squares of the reciprocal natural frequencies of oscillation. Components of the eigenvector corresponding to a given eigenvalue are the magnitudes of the loop currents at the corresponding frequency.

Since lossless networks cannot be built in practice, let us investigate the effect of losses in the network of Figure A-1. If a small resistance is connected in series with every inductance such that the frequency of oscillation is $\omega_d = \sqrt{\omega_o^2 - \alpha^2}$, where α is the real part of the coordinates of the poles of the network, the error in using the damped natural frequencies ω_d in place of the undamped frequencies may be calculated. The percentage error of determining the eigenvalues is given in Equation A.6 expressed in terms of the Q of the resonant circuits.

$$\% \text{ error in eigenvalues} = \frac{100}{(2Q)^2 - 1} \qquad (A.6)$$

Even for a lossy network having a Q of 10, the error is only 0.25%. We may thus draw the conclusion that network losses don't seriously affect the accuracy of the eigenvalues.

The eigenvalues may be obtained by spectrum analysis of any of the loop currents. This is readily accomplished by feeding the voltage across any of the series resistances into a tunable narrow-band filter whose tuning frequencies corresponding to peak outputs yield the eigenvalues. The corresponding eigenvector may be obtained by sampling the output amplitudes of synchronously tuned narrow filters connected to measure each loop current. The samples are taken when local peak outputs with tuning are observed.

The size of the matrix solved by the preceding methods may be made arbitrarily large. The reader can readily verify that if the matrix whose eigenvalues and vectors we wish to compute is N x N, then the network topology has to consist of N nodes which are connected to each other and to ground by series LC networks as illustrated by Figure A-3 for N= 3.

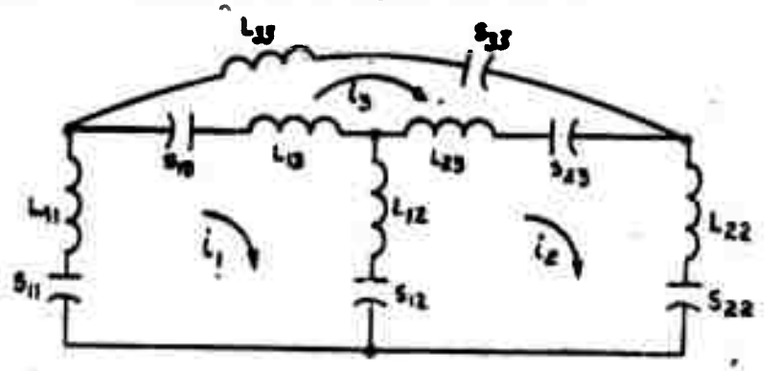


Figure A-3. Generalization of Eigenvalue Problem

APPENDIX B. ON PATTERN RECOGNITION

In this Appendix we briefly outline a number of approaches to the subject which have been taken by other people.* There is a thread of similarity among all of them. Yet they all differ, and some widely, for they each ask their questions from separate frames of reference. And as Susanne Langer and Sir Arthur Eddington have suggested — the way we ask a question determines or frames our answer. The ichthyologist who sets out to study the ocean life with a net having two inch openings comes to the "remarkable" conclusion that there are no fish smaller than two inches. So too, each pattern recognition technique is limited. A model chosen to simulate the human nervous system is limited by the "designer's" understanding of the nervous system and also by the fact that the biological system may not be the best one for the problem under examination.

Pattern recognition is the examination and classification of an object as a member of one of a number of categories, rather than its unique identification. The recognition process is information-destroying, and only the invariant properties or characteristics of the pattern (or patterns) of interest convey significant information for classification. Of course, in order to perform a classification, it is necessary to know the pattern-characteristic properties. Perhaps these properties can be chosen through human ingenuity and study; and this, in fact, has been the basis for most character recognition and speech recognition systems now in vogue. But the more basic approach to the problem attempts to develop a technique

* In the preceding sections of this report, methods are based on the doctoral thesis work of George Sebestyen.

whereby supervised examination of known samples of the various categories permits a computer to "learn" what the parameters are which are significant to each category. The approach taken in this report is of the latter type.

Pattern recognition is closely related to problem solving. Problem solving consists of examining a given set of elements and finding a member of a subset having specified properties. Use is made of processes which find possible solutions, and of other processes which determine whether a proposed solution is in fact a solution. Use is made of principles or devices called heuristics, which contribute to the simplification of the search for solutions.

The actual measure of the merit of any particular pattern recognition technique is in the success of the decisions. All objects to be classified can be represented in a space — perhaps an n-dimensional vector space. If the objects can be associated with particular categories, then by definition there is a decision rule (perhaps a rule defined on a point set) for perfectly classifying every object, given its location in the vector space (assuming no noise, of course). The "learning" task is to find the decision rule from examination of a finite number of known examples. Clearly, the goal is a technique which can find a good decision rule rapidly as the number of samples is increased. All of the techniques make use of heuristics. These are aids or constraints which limit the number of solutions under consideration. These heuristics are chosen on the basis of the designer's experience and/or imagination. They could be suggested by a sophisticated computer. Often the heuristics chosen tend to degrade the technique by hiding the best solutions. This simply suggests that a good

heuristic can be a big help whereas a poorly chosen one can eliminate the proper solutions altogether. The differences among the various techniques lie, therefore, in the manner in which each leads toward development (or discovery) of a decision rule.

We have heard the term artificial intelligence. This is generally applied in study of machines which are told how to learn, but not how, specifically, to perform a task. We might further note the subdivision in pattern recognition into cases where the machine is told how to find what makes the objects of a particular category equivalent, and cases where the machine does not have this information. Thus we see that the class of pattern recognition techniques includes machines which are told just how to identify objects (from given truth tables), machines which are supplied with helpful hints (heuristics) on how to learn to classify objects, and machines which are told nothing and are allowed to organize themselves by trial-and-error and through supervised learning. In the extreme, of course, this latter type of machine has a greater hardship imposed upon it than does the learning human mind.

Supervised learning consists in presenting known samples to the machine and telling it to which category the object belongs. This is made known to the machine through a process of reward-and-punishment. The supervisor naturally imposes his own notions on just what constitutes the pattern. From examination of samples used in character recognition studies one immediately realizes that a particular letter is classed in a particular category only because the supervisor (in this case, perhaps the one who drew the character) chooses to classify this character in this manner. To any other human being this character might be confused with another or it might even be unintelligible.

Many researchers have approached the subject through study and subsequent attempted simulation of the human nervous system. Still other have felt that the best approach is to start fresh by attempting to devise a model to solve the problem directly, without recourse to biological justifications. It is interesting to note that the models developed from these markedly different approaches have stron similarities with each other. This could lead to the encouraging conclusion that progress has been good since different approaches lead to similar solutions. More likely, however, the conclusions might be that human researchers are again caught in that common quagmire wherein they persist in asking their questions within the old framework, and then are surprised when they get the same old answers. The observer from the next dimension chuckles as he watches our dilemma, just as did the three-dimensional visitor in the Flatland of Edwin Abbott. A completely new approach to the problem could be most refreshing, but of course its arrival is not now predictable.

The bibliography at the end of this report can serve as a survey of pattern recognition. In reviewing the work done in pattern recognition let us first examine the studies of Newell, Simon, and Shaw, at Rand, on the processes of creative thinking and applications to a Logic Theorem Proving machine and a chess playing machine. Strictly speaking, the work is not called pattern recognition, but the ideas are interesting and are here presented in more than just a passing manner. Creative activity is a special class of problem-solving activity characterized by novelty, unconventionality, persistence, and difficulty in problem formulation. The Logic Theorist (a computer program, and possibly a machine) attempts to prove theorems (handed to it) of the type found in Principia Mathematica, and in proving the theorems it then conjectures

and proves new theorems on which the original proofs depend.

We earlier defined problem solving and discussed the processes of generation of possible solutions and od determining whether a proposed solution is in fact a solution. Apparently, for large difficult problems, there may be large correlation between creativity and use of trial-and-error generators.

The Logic Theorist operates on only a restricted set of proofs, and tests these. The restrictions are on the number of logic expressions, number of symbols in each expression, and number of different kinds of symbols used. We may further restrict the algorithm by only considering sequences that are valid proofs — i.e., whose initial expressions are axioms, and each of whose expressions is derived from prior ones by valid rules of inference. Now, one approach could be to generate first the shortest proofs, and then longer ones by applying the rules of inference (in all possible ways) to the former (shorter) proofs. This is working forward. Actually, the Logic Theorist works backwards. The Logic Theorist generates proofs which contain the desired expression (theorem) for the final one, and logical expressions (obtained from logical inference) for the preceding ones. When a proof appears whose initial expressions are theorems, we have found the desired proof.

We can specify a solution by either specification by state description or by specification by process description. For example, in logic we can write out an expression in the usual way, or we can give a sequence of operations on the axioms (a proof) that will produce it.

We earlier defined the term heuristic to denote a principle or device that reduces, on the average, the search required to reach a solution. Many

of the restrictions mentioned earlier are heuristics. Heuristics are processes to select correctly a very small part of the total problem—solving maze for exploration. Most heuristics depend on a strategy that modifies subsequent search as a function of information obtained in previous search. Note that algorithms are foolproof heuristics. Other heuristics are the following. In logic-proofs, apply an operator if this results in an expression which more closely resembles the final expression than did the previous one. Another heuristic is to set up sub-tasks. A graphic example of such problem factorization is shown in the case of a safe with a lock having 10 independent dials numbered from 00 to 99. Random twirling of dials would require, on the average, $1/2 \times 10^{20}$ trials to open the lock. If the lock is defective and there is a faint click each time any dial is turned to its correct setting, then on the average only 500 trials are required to open the lock (50 trials for each dial). We might also note that "insight" into the problem structure is actually the acquisition of an additional heuristic. It may be of further interest to observe here that our pre-processing of data makes use of many heuristics.

These ideas on thinking have been presented for their general interest in the field. Some specific pattern recognition techniques follow.

Pandemonium. The model of Pandemonium originated by O. Selfridge consists of four stages of devices. The first stage consists of data collection (and display) devices, the second of computational devices, the third of cognitive devices, and the fourth of a decision unit which selects the cognitive device having the largest output. The cognitive devices are each associated with a particular category (pattern) in the classification problem being solved. These latter devices each measure (in some sense) the similarity between the particular pattern the cognitive device represents and the as-yet-unclassified input. Each cognitive device has an output proportional to the amount of the aforementioned similarity. The unknown inputs are introduced to the data collection devices (through which the physical word is represented), and on which the computational devices operate. These latter extract the various "features" of the pattern and give an output proportional to the amount of the respective features. Between each computational device and each cognitive device there is a weighting network, the value (weighting) of which is determined during the process of supervised learning. The weightings emphasize the features most significant for each pattern, and the process of developing the weightings is known as "feature weighting". The Pandemonium is programmed to adjust its weighting to minimize the output of the appropriate cognitive device.

Perceptron. The Perceptron is a generic name for a family of pattern recognition machines (oringinally proposed by Frank Rosenblatt) that operate on principles not unlike those believed used in the human brain. The perceptron consists of sensory units, associative units (each one is an effect a variable memory unit in a large switchboard), and response units. There is

a response unit corresponding to each stimulus class (category), and each gives an output proportional to the similarity between the pattern which it represents and the unknown input. A decision device selects the largest output. The outputs of the response units are fed back to weighting networks, located between the associative units and the response units, in such a manner that the connections contributing to the correct output are strengthened and the connections contributing to incorrect outputs are inhibited. As the stimulii are sequentially applied to the sensory units (during the period of supervised learning), the weightings are readjusted until the Perceptron approaches its asymptotic learning capability.

In elementary models of the Perceptron the connections are linear, and the associative units are simple discrete representations of a neuron. In more sophisticated models, the connections may be defined with non-linear functions, and the associative units have more complex properties. For example, in more advanced concepts of the Perceptron, the associative unit itself adjusts its firing threshold and the value of its output in the course of the learning process.

Other Related Approaches. The Perceptron and Pandemonium are among the models more widely known (by name) in this country. But there are other approaches which merit as much examination as these two. Several people in England have introduced their own models. Chapman proposes a model in which the memory cell has certain special properties. Everytime a cell fires under stimulation it modifies the structure of the triggering cells differently from the passive ones. Uttley has done work on conditional probability computers and on methods of classification, following the study of the human nervous system. Models follow from this. Taylor has done similar work.

In this country, Mattson has studied the classification problem with a model in which he represents objects in an n-dimensional binary space. He then divides the space into category regions with a number of hyperplanes. The "learning" process consists in best locating the hyperplanes, that is, in finding the coefficients defining each plane. The evaluation is made with logical networks. Simple models have been constructed. Stanford University is looking at the problem with a similar viewpoint.

Nerve Nets. McCulloch and Pitts have approached the problem by concentrating their efforts on development of a model for a neuron. They then study the properties of the different (nerve) nets which can be synthesized with these neuron models. Whereas some people begin with a system, perhaps resembling the biological model, and try to learn the requirements for its structure, McCulloch and Pitts start with the basic element and study its applications in large systems. They have directed their more recent efforts to further refinement of the basic nerve model.

Character and Speech Recognition. Most studies of character and speech recognition have not included real learning techniques, but have rather depended upon looking for particular features in a pattern as suggested by the ingenuity of the engineer. Mattson and Rosenblatt have applied their techniques to character recognition, and other work is being done at Stanford Research Institute, Lincoln Labs, Bell Telephone Labs, and at other organizations. One procedure which appears to be widespread is the following: the character is quantized in its two-dimensional display; noise is removed by an operation of local averaging, i.e., of representing a box (quantum) by the average of itself and all immediately adjacent boxes; the character line width is standardized; certain character features are then extracted and the

recognition is based upon the existence of certain features. It is in the feature selection that the designer's ingenuity is manifested. These features may include different line orientations and straight line intersections, such as a T, inverted T, a V, a slant, etc. In some techniques the original character is quantized and then converted to binary form, and all subsequent operations are logical (performed by a digital computer). In most of the techniques location of character and size variations are special problems.

The techniques of speech recognition are nicely described in C. Cherry's book "On Human Communications", and in Bell Telephone Laboratories monographs. Speech is represented in a two-dimensional time-frequency array which is obtained by passing the speech through an array of staggered narrow-band filters (in a device known as a VOCODER). Some particular speech recognition techniques attempt to extract properties from these arrays, such properties including formant frequencies, etc. Special problems arise from the wide ranges in pitch and word duration among speakers, in addition to the other many subtleties of the spoken word.

Language Translation

Strictly speaking, language translation is a problem in more specific identification rather than in pattern recognition. However, many of the translation processes do involve a search for patterns. For example, any particular word sequence must be examined to see whether it has the pattern of a grammatically correct sentence, i.e., is the relationship of verb, noun, adjective, etc. consistent with the rules of syntax?.

Turing Machines

One of the early classical approaches to examination of the potentialities of machine learning was introduced by Turing. Claude Shannon presents

a description of a Universal Turing Machine in "Automata Studies" (Princeton University Press, 1956), and his introductory description is here given exactly.

> "In a well-known paper[1], A.M. Turing defined a class of computing machines now known as Turing machines. We may think of a Turing machine as composed of three parts — a control element, a reading and writing head, and an infinite tape. The tape is divided into a sequence of squares each of which can carry any symbol from a finite alphabet. The reading head will at a given time scan one square of the tape. It can read the symbol written there and, under directions from the control element, can write a new symbol and also move one square to the right or left. The control element is a device with a finite number of internal "states". At a given time, the next operation of the machine is determined by the current state of the control element and the symbol that is being read by the reading head. This operation will consist of three parts; first the printing of a new symbol in the present square (which may, of course, be the same as the symbol just read); second, the passage of the control element to a new state (which may also be the same as the previous state); and third, movement of the reading head one square to the right or left.
>
> "In operation, some finite portion of the tape is prepared with a starting sequence of symbols, the remainder of the tape being left blank (i.e., registering a particular "blank" symbol). The reading head is placed at a particular starting square and the machine

proceeds to compute in accordance with its rules of operation. In Turing's original formulation, alternate squares were reserved for the final answer, the others being used for intermediate calculations. This and other details of the original definition have been varied in later formulations of the theory.

"Turing showed that it is possible to design a universal machine which will be able to act like any particular Turing machine when supplied with a description of that machine. The description is placed on the tape of the universal machine in accordance with a certain code, as is also the starting sequence of the particular machine. The universal machine then imitates the operation of the particular machine".

There are many detailed variations in the approaches outlined above. Another way of looking at the subject is contributed by Bellman and Kalaba, and is built upon the dynamic programming techniques which these men have developed. Further information on pattern recognition may be outlined from the bibliography.

APPENDIX C

As experimental verification of the technique, the methods of Section 2 and 3 were applied to the machine-learned recognition of spoken numerals. The sequence of labeled events is a large set of numerals, spoken by different individuals, where each spoken word is labeled by one of the ten numerals it represents. An unlabeled spoken word is recognized as a specific numeral through its comparison to each of the ten categories by the functions developed from the labeled examples. The ten categories of spoken numerals "0" (zero) through "9" were represented by 400 different utterances made by ten male speakers. The ten male speakers have regional accents drawn from the north-east corner of the United States. No attempt was made to otherwise control the selection of speakers or their rate of speech.

The model of the physical world considered adequate to represent the speech events is obtained through use of an 18-channel Vocoder. The Vocoder is a set of 18 stagger-tuned narrow band-pass filters which print out the "instantaneous" frequency spectrum of the speech event as a function of time. This is shown in Figure C-1, where frequency is plotted vertically, time horizontally, and the intensity of the spectrum at a given frequency and time is proportional to the grey level of the sonograph recording at the corresponding time-frequency point. The numerical print-out of Figure C-1 is obtained by digitizing the above sonograph records into 18 frequency channels each sampled at the rate of 20 m sec/sample. Note that the samples are orthogonal by construction because they represent waveforms that are disjoint either in frequency or time. The resulting cell structure in the time-frequency plane represents a one-second long speech event as a vector

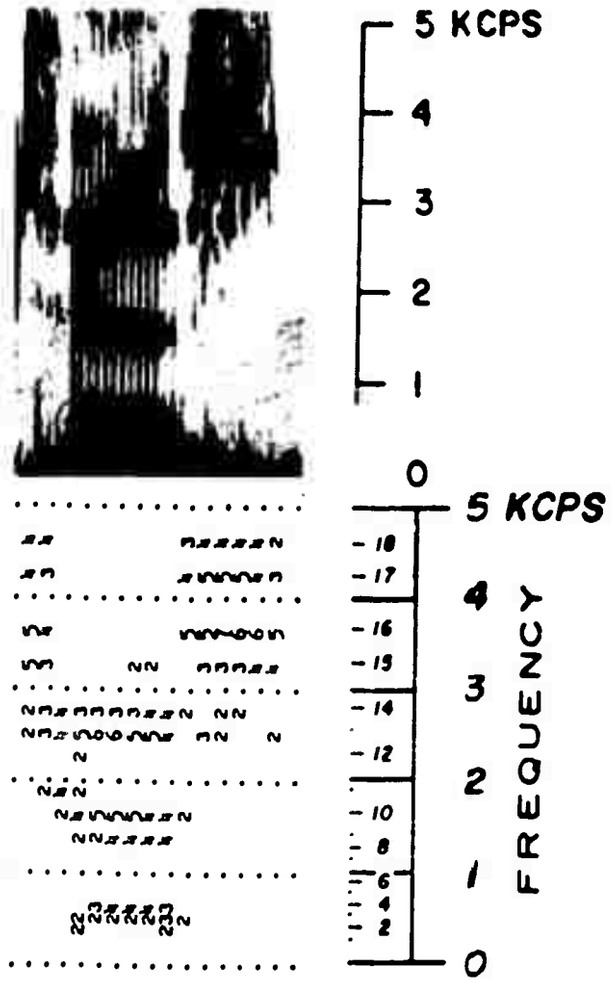


Figure C-1. The Spoken Word "Test"

in a 900-dimensional space. Each dimension corresponds to a possible cell location and the coordinate value of a dimension is the intensity of the corresponding cell. In the figure a three digit binary number (eight levels) represents the sonograph intensity, after the instantaneous total speech intensity is normalized by the action of a fast age.

A computer was programmed to implement the theory; and an increasing number of spoken numerals were sequentially introduced from which the computer constructed the optimum metrics. A different metric was developed to measure similarity to each of the ten categories. Typical results of the learning process are shown in Figure C-2. This figure contains four confusion matrices constructed for the cases where numeral recognition was learned from 3, 4, 7, and 9 examples of each of the ten categories of digits. The ordinate of a cell in the matrix signifies the digit which is spoken, the abscissa denotes the decision of the machine, and the number in the cell states the number of instances in which the stated decision was made. The number 1 in row 6 and column 8 of Figure C-2c, for example, denotes the fact that in one instance a spoken digit 6 was recognized as an 8. Note that the error rate decreases as the number of known examples of categories is increased. For the 9 examples per category no errors were made. This result is particularly interesting in view of the fact that the spoken digits which were tested were spoken by persons not included among those whose words were used as examples.

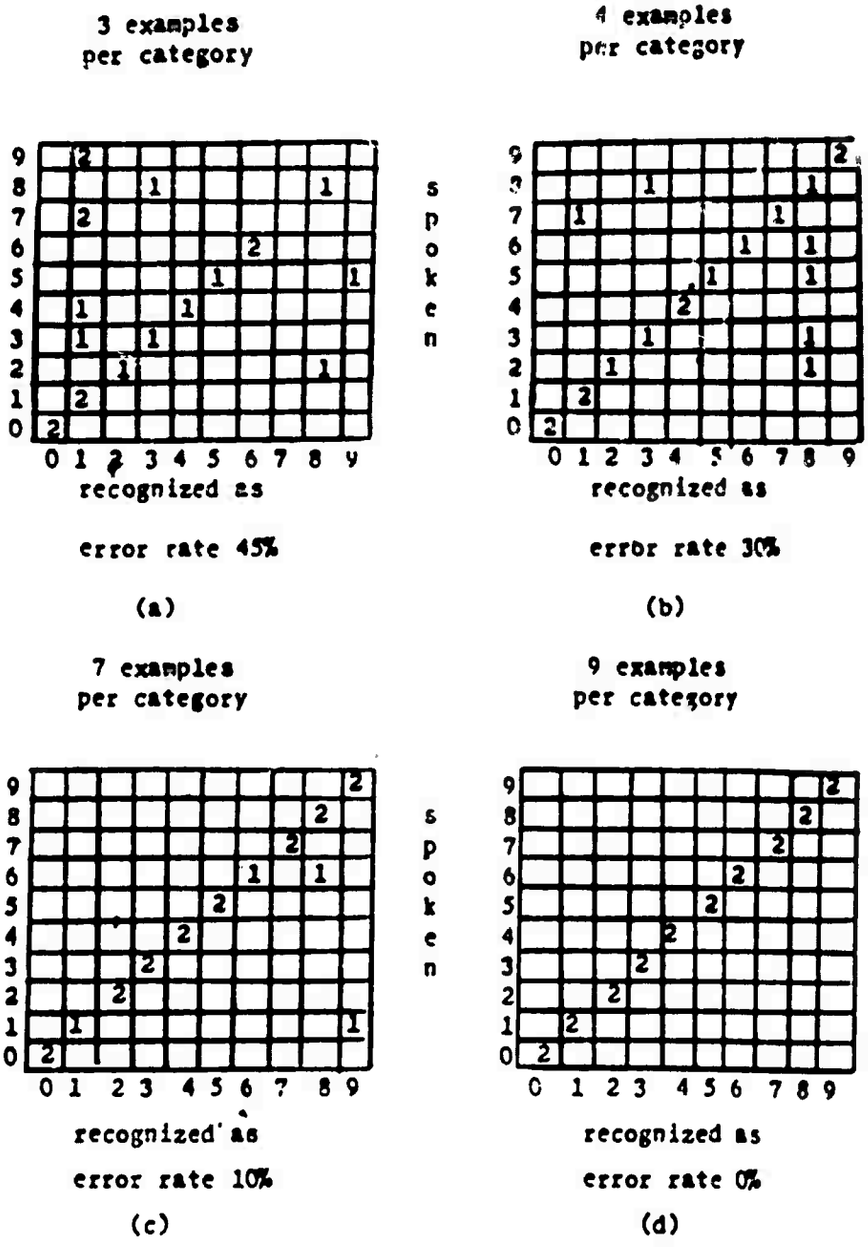


Figure C-2. Confusion Matrices Illustrating Learning Numeral Recognition

UNCLASSIFIED